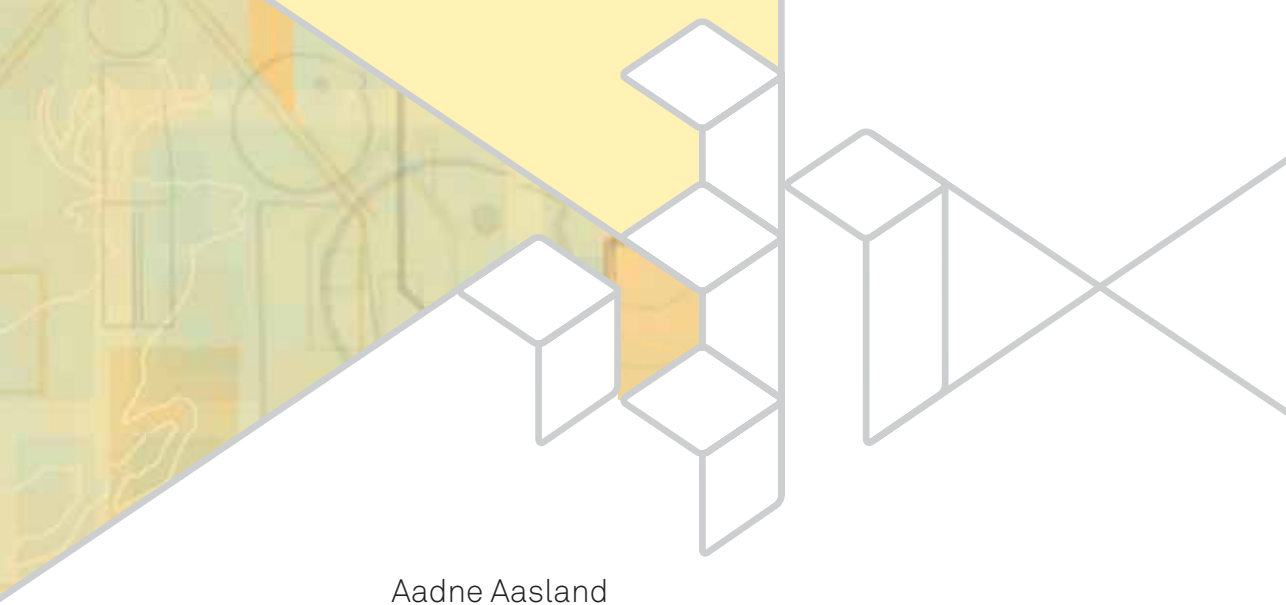




Aadne Aasland

A User Manual for SPSS Analysis:

CNAS 2008 Survey Data



NIBR

Norwegian Institute for Urban and Regional Research

Aadne Aasland

A User Manual for SPSS Analysis:

CNAS 2008 Survey Data

Table of Contents

Preface	2
1 Introduction to the CNAS 2008 survey data	3
2 Types of data analysis	6
3 Preparing the data for analysis: Exploratory analysis and data cleaning	7
3.1 Distribution of the data	7
3.2 Cleaning the data	8
3.3 Weights	9
4 Univariate analysis	13
4.1 The distribution	13
4.2 Central tendency	16
4.3 Dispersion	17
5 Comparing groups: Bivariate analysis	19
5.1 Bivariate measures of association and significance tests	23
6 Creating additive indexes	29
7 Multivariate analysis	35
7.1 Multiple linear regression	35
7.2 Logistic regression	39
8 Presenting your findings – making tables and graphs	46

Preface

In the winter and spring of 2008 the Centre for Nepal and Asian Studies (CNAS), Tribhuvan University and Shtrii Shakti (S2), in close collaboration with the Norwegian Institute for Urban and Regional Research, conducted two large-scale household surveys as part of a 3-year project on social inclusion and exclusion in Nepal. The aim of this manual is to demonstrate step-by-step a variety of the techniques that can be effectively applied for data analysis of the complex survey data. There are examples of basic analysis techniques as well as more advanced techniques that enable the researcher to answer complex questions that cannot be answered through simpler forms of analysis.

It is our hope that the manual will be useful for students of quantitative methodology in Nepal, and especially those who engage with the topic of inclusion and exclusion. A training course on quantitative survey analysis was carried out in Kathmandu in November 2008, and much of the manual is based on input before, during and after this course. It is meant to be very practically oriented with a focus on applied methodology and analysis.

The reader should be familiar with basic statistics, or be aided by statistics handbooks during the work with this manual. Also, the manual requires access to a survey data set. We decided to use the CNAS data set which is the most comprehensive in terms of dimensions of exclusion. This data set can be provided free of charge to enrolled students and researchers, by approaching CNAS.

We would like to thank all those in CNAS and S2 who have contributed to the two surveys and the people they have hired to participate – in sample design, data collection, data entry and data cleaning. Particularly we wish to thank project coordinator, Professor Dilli Ram Dahal of CNAS. Furthermore, Associate Professor Bidhan Acharya, Population Studies, Tribhuvan University has been in charge of the sampling design used for the CNAS survey and has prepared the data for analysis. We also thank Berit Willumsen for help in preparing the manuscript for publication. Finally, we are very grateful to the Ministry of Foreign Affairs of Norway for its generous financial support.

Oslo, September 2009

Marit Haug

Research Director
Project Leader

1 Introduction to the CNAS 2008 survey data

Data analysis will never provide good results unless the data are of good quality. Therefore, already in the preparation phase of a project great care needs to be taken to use operational definitions that are valid and reliable measures of concepts.¹

This manual is based on an existing data set from a survey on social exclusion and inclusion in Nepal. Preparations for data analysis starts already in the planning phase of a survey, with questionnaire design and procedures for sampling. As this manual is primarily concerned with data analysis techniques, topics such as questionnaire design, sampling and other preparatory work are not treated here. Nevertheless, one can hardly overestimate the importance of these preparatory phases.

The appropriate methods of data analysis are determined by your data types and variables of interest, the actual distribution of the variables, and the number of cases. In the case of the CNAS data set, these parameters are given for those who wish to analyse the data.

It is important to have an initial understanding of the survey data set that is used for this manual. The CNAS data set was collected in four districts of Nepal: Dhanusa, Sindhupawlhuk, Surkhet and Banke. In each district the aim was to have 600 respondents, (but 1,200 in Dhanusa with two target groups). Of these 400 were to be selected from the target groups (Tarai Dalits and Yadavs in Dhanusa, Tamangs in Sindhupawlhuk, Hill Dalits in Surkhet, and Muslims in Banke). The remaining 200 were to be selected among the non-target groups (general population). In each district a stratification took place whereby 20 research sites were selected. For selection procedures and overall survey methodology, see the CNAS project report.²

This manual requires some familiarity with SPSS for Windows. Thus, it will not cover the more general procedures in SPSS. There are a number of SPSS courses available for students and researchers to familiarize themselves with the programme, and it is recommended that some basic skills are already developed before getting to work on the CNAS data, which is a rather complex data file.

When you receive the CNAS data set, the following preparatory work has already taken place:

¹ A measure is **valid** if it actually measures the concept we are attempting to measure. It is **reliable** if it consistently produces the same result.

² Forthcoming in the autumn 2009.

- - Data have been entered into a data file in SPSS for Windows with cases (the respondents) in rows, and with variables (based on survey questions) in columns. This is what you find if you look at the data file in Data view. In the Variable view you find all the variables in Columns and some characteristics of each variable (which you are allowed to change) in columns.
- - Some key variables have been recoded or computed into new variables that were not originally in the questionnaire based on combining responses from two or more variables or regrouping responses on one variable. The variable and value labels should explain these new variables. For example: age at birth has been recoded into age groups.
- - Missing values and variable types (see later) have been assigned to all variables where relevant.

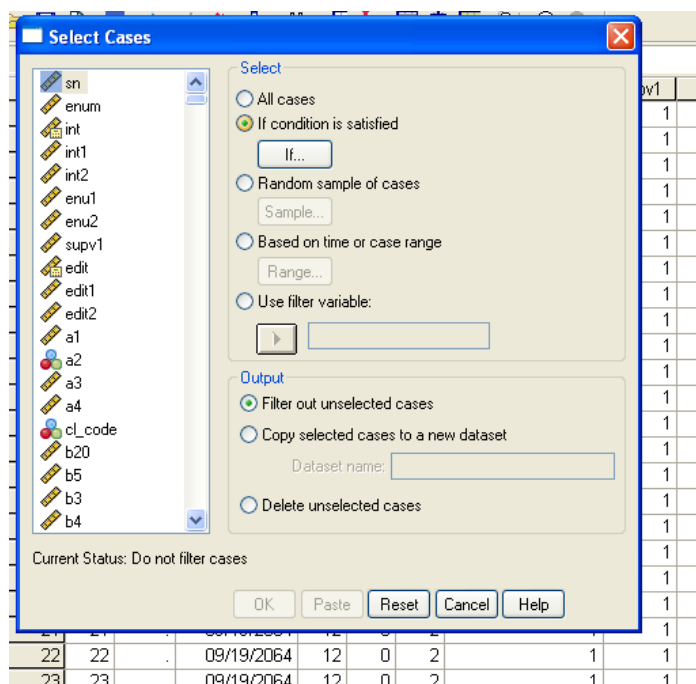
Before using the data, you should save it as your own working data file, in order to preserve the original data. In case you make an error, you can then revert to the original data file. It is very often useful to save all the syntax you use for computing new variables, then you can simply run the syntax file again if your working data file suddenly contains errors that you are not able to remove. You do this by saving the data with a new name that is easy to identify, e.g. *Save as CNAS_aaa1.sav*. You can save as many data files as you wish (but of course they make up some space on your hard drive). You can also put the date in the name of the data file so that it is easy to see when it was created, e.g. *CNAS_220909.sav*.

You will need a CNAS survey questionnaire to analyse the data, so that you can see the wording of each question. The variable names usually reflect the code for each variable in the questionnaire. Thus, the questionnaire contains sections from A to S, in addition to some administrative variables, most of which you find at the beginning of the data file. The data are normally sorted according to the letters in the alphabet, but you can also sort them according to when they appeared in the data file.

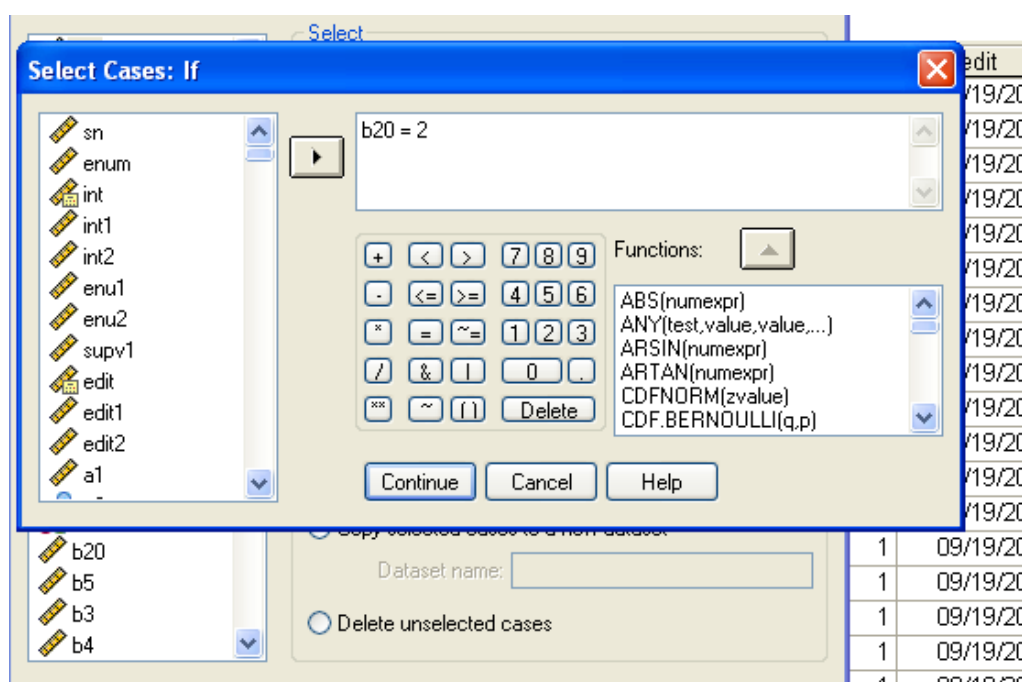
The CNAS survey data enable three types of analysis:

1. Analysis on all household members (mostly from section B).
2. Analysis on the household as such (A section, most of C section, much of D section, etc.)
3. For one randomly selected individual in the household (most of the remaining sections)

It is very important to note that the data file contains data on each individual in the household. Thus, as it is, it is mostly suited for analysis in section B. **If you wish to carry out analysis on the randomly selected individual (the respondent) you should do analysis only in cases where B20 (Survey status) = 2 (Selected respondent)** where all the respondent and input is recorded. This is also the case if you wish to Household level and individual level. You do this by opting for Select Cases under Data in the scroll-down menu, tick for If condition is satisfied



Click If... under If condition is satisfied.



In the empty window, write `b20 = 2`, and click continue.

The first window comes back, and click OK. For the subsequent analysis you will only analyse cases for respondents (or households).

If you wish to do analysis only for one district or only for one ethnic group, you use the same procedure. You can combine by writing e.g.

`B20 = 2 AND district = 1`.

2 Types of data analysis

It is common to differentiate between three different types of data analysis, and we will go through all the three in the next chapters:

Exploratory Data Analysis

Exploratory data analysis is used to quickly produce and visualise simple summaries of data sets. We use exploratory data analysis mostly for arranging the data for further analysis.

Descriptive Data Analysis

Descriptive statistics tell us how the data look, and what the relationships are between the different variables in the data set. We perform descriptive data analysis to present quantitative descriptions in a manageable form.

It should be noted that every time we try to describe a large set of observations with a single indicator, we run the risk of distorting the original data or losing important detail. However, given these limitations, descriptive statistics provide a powerful summary that may enable comparisons across groups of people or other units.

Inferential Statistics

Inferential statistics test hypotheses about the data that makes it possible to generalize beyond our data set. We will come back to inferential statistics in the section below on comparing groups.

It is also common to differentiate between the three following types of statistical analyses:

1. *Univariate*- when one variable is analyzed
2. *Bivariate*- analysis of two variables
3. *Multivariate*- analysis of three or more variables

In the following we will start by discussing the main principles of exploratory data analysis. It will be followed by examples of univariate, bivariate and multivariate analysis techniques, involving both descriptive data analysis and inferential statistics.

3 Preparing the data for analysis: Exploratory analysis and data cleaning

The first task once the data is collected and entered is to ask: "What do the data look like?".

Exploratory data analysis uses numerical and graphical methods to display important features of the data set. Such exploratory data analysis helps us to highlight general features of the data and thereby direct our further analyses. In addition, exploratory data analysis is used to highlight problem areas in the data. One should particularly ask the following:

- What do the distributions look like for key variables?
- To what extent do the data need cleaning for consistency?
- Should outliers (values that are far from the other values in the distribution) be included or excluded in the analyses?
- Are there many cases and variables with missing data, and how should such missing data be handled?

3.1 Distribution of the data

First we go through the data file and investigate the "shape" of the data. Where do most of the values lie? Are they clumped around a central value, and if so, are there roughly as many above this value as below it? We look at the distribution for each variable to determine which analyses would be most appropriate. Types of analyses are also determined by the types of the variables (nominal, ordinal or scale levels).

In SPSS you can specify the level of measurement as scale (numeric data on an interval or ratio scale), ordinal, or nominal.

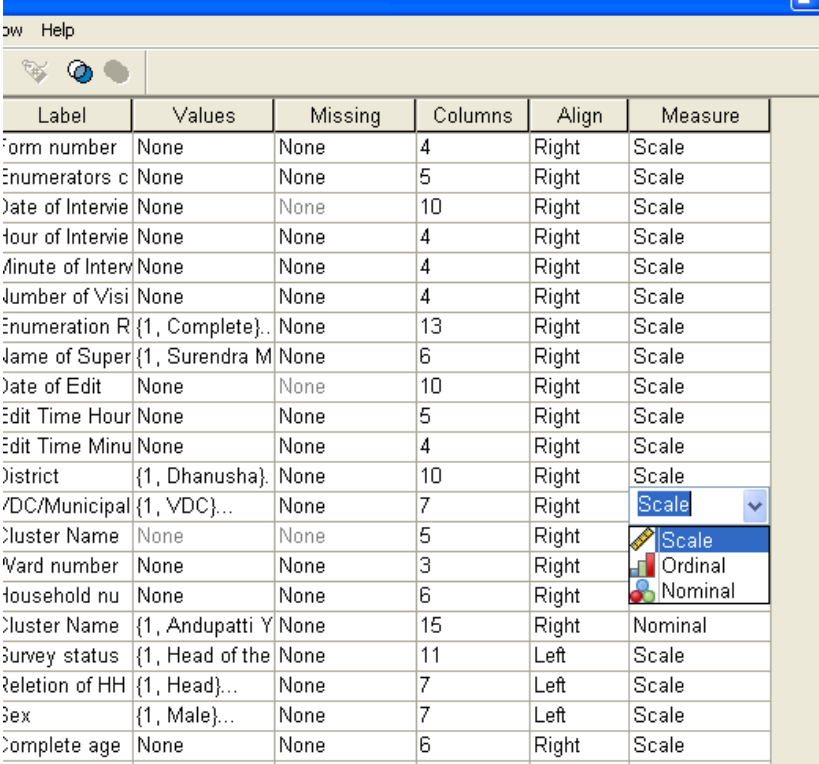
A variable can be defined as *nominal* when its values represent categories with no intrinsic ranking. Examples of nominal variables in our data set include VDC/municipality (A2), sex (B3), ethnicity (B6) and religious affiliation (B7).

A variable can be defined as *ordinal* when the values represent categories with some intrinsic ranking; for example, levels of satisfaction from highly dissatisfied to highly satisfied. Examples of ordinal variables in the data set include attitude scores, such as comparing income situation today with that of 25 years ago (highly improved, somewhat improved, ... etc.) (D15), and how person the respondent is to be a person of his/her caste or ethnicity (very proud, somewhat proud, ... etc.) (O15).

A variable can be defined as *scale* when the values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables from the survey include age in years (B4) and income in Nepali rupies (B14).

Exercise: Go through the data file and check the variables. Define them according to their measurement level: Nominal, ordinal or scale. Save the file in a new name, and use it as your new working file.

Hint: go to the variable view of your data file. Define measurement level in the box to the right (under Measure).



Label	Values	Missing	Columns	Align	Measure
Form number	None	None	4	Right	Scale
Enumerators c	None	None	5	Right	Scale
Date of Intervie	None	None	10	Right	Scale
Hour of Intervie	None	None	4	Right	Scale
Minute of Interv	None	None	4	Right	Scale
Number of Visi	None	None	4	Right	Scale
Enumeration R	{1, Complete}..	None	13	Right	Scale
Name of Super	{1, Surendra M	None	6	Right	Scale
Date of Edit	None	None	10	Right	Scale
Edit Time Hour	None	None	5	Right	Scale
Edit Time Minu	None	None	4	Right	Scale
District	{1, Dhanusha}.	None	10	Right	Scale
VDC/Municipal	{1, VDC}...	None	7	Right	Scale
Cluster Name	None	None	5	Right	Scale
Vard number	None	None	3	Right	Scale
Household nu	None	None	6	Right	Scale
Cluster Name	{1, Andupatti Y	None	15	Right	Nominal
Survey status	{1, Head of the	None	11	Left	Scale
Relation of HH	{1, Head}...	None	7	Left	Scale
Sex	{1, Male}...	None	7	Left	Scale
Complete age	None	None	6	Right	Scale

3.2 Cleaning the data

During the exploratory data analyses we assess the need to clean our data. Data cleaning is extremely important, and especially when the data collection method allows inconsistencies. All data cleaning work should be carefully documented and available in a report. Data cleaning includes, among others, the following

- Removal of invalid, impossible, or extreme values. Such data may be removed from the dataset and recoded as missing values. Unusual values may be out of range, physically impossible (a person of 149 years), unrealistic (an income of 10000000000 Nepali rupies per month), etc. Outliers might also be marked for exclusion for the purpose of certain analyses.

- Labeling missing values: It may be necessary to label each missing value with the reason it is considered missing in order to guarantee accurate bases for analysis.

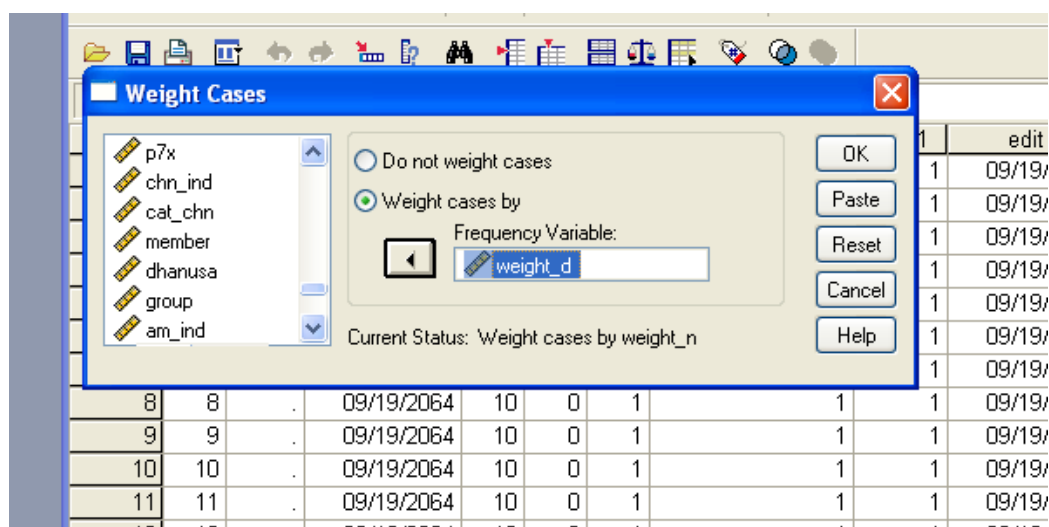
The data that you have received should be cleaned, but sometimes we discover certain inconsistencies during data analysis. One should then perform the appropriate cleaning. Serious inconsistencies that are found should be reported to CNAS.

In a survey, missing values correspond to skipped questions or impossible options. A discussion in the research team should take place in determining how missing values should be handled. In some cases, missing values might be perfectly normal (e.g. the variable "How many livestock are there with your family with different category" - C12a to C12o - should only be answered by those households who in C11 said that their families keep livestock). However, in some cases missing values for important variables might exclude a record from certain analyses. Sometimes it is appropriate to place normalized values in place of missing values. We will come back to this when we go through how to compute additive indices below.

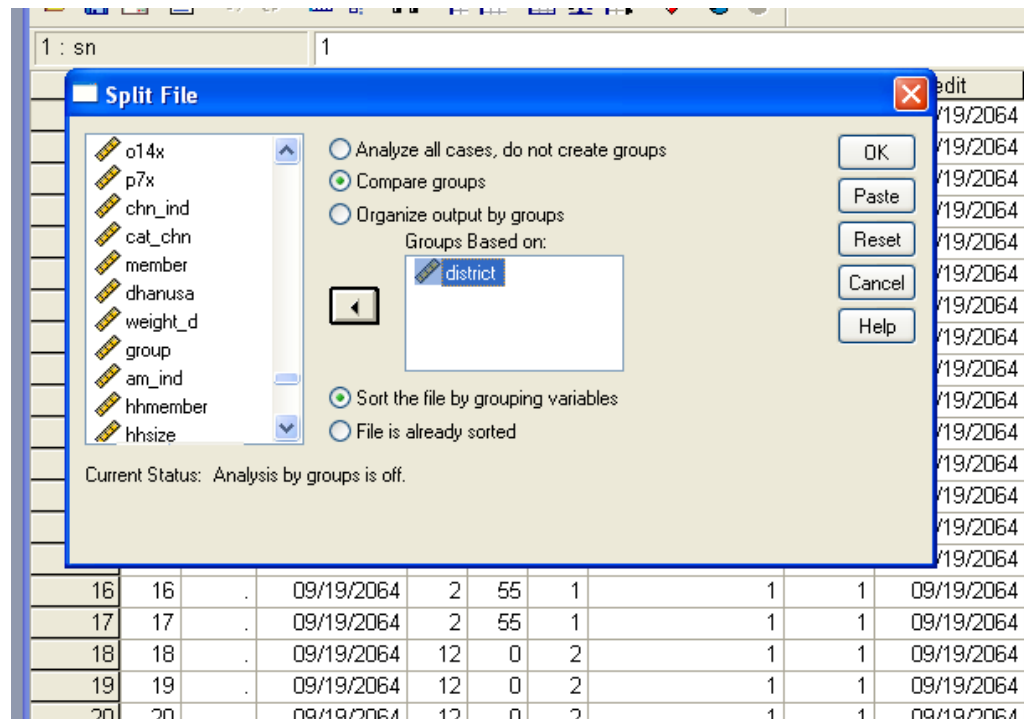
3.3 Weights

Since the number of certain target groups make up a larger share of the sample than their share in the population, we get biased results unless we weight for such discrepancies. Therefore, based on population data in the four selected districts, those groups that are over-represented (Tarai Dalits and Yadavs in Dhanusa, Tamangs in Sindhupawlchuk, Hill Dalits in Surkhet and Muslims in Banke) are given a weight (the variable is called `weight_d`) so that their proportion in the analysis reflects their proportion in the population. The same goes for all other groups. In order to apply these weights do the following:

1. When in the Data window, choose Data and Weights, select `weight_d`.



However, note that the data are not representative of Nepal as such. To get correct results for each district, one should split file by district and treat each district separately.



Before weighting, we had the following distribution of respondents belonging to target and non-target groups in each district:

target1 Target Population							
district Survey district				Frequency	Percent	Valid Percent	Cumulative Percent
1,00 Dhanusa	Valid	1	Selected Ethnic Group	817	68,8	68,8	68,8
		2	All Others	370	31,2	31,2	100,0
			Total	1187	100,0	100,0	
2,00 Sindhupawljchuk	Valid	1	Selected Ethnic Group	360	65,8	65,8	65,8
		2	All Others	187	34,2	34,2	100,0
			Total	547	100,0	100,0	
3,00 Surkhet	Valid	1	Selected Ethnic Group	405	68,5	68,5	68,5
		2	All Others	186	31,5	31,5	100,0
			Total	591	100,0	100,0	
4,00 Banke	Valid	1	Selected Ethnic Group	393	69,6	69,6	69,6
		2	All Others	172	30,4	30,4	100,0
			Total	565	100,0	100,0	

However, after weighting we get the following distribution:

target1 Target Population

district	Survey district		Frequency	Percent	Valid Percent	Cumulative Percent
1,00	Dhanusa	Valid				
		1 Selected Ethnic Group	343	29,6	29,6	29,6
		2 All Others	813	70,4	70,4	100,0
		Total	1156	100,0	100,0	
2,00	Sindhupawichuk	Valid				
		1 Selected Ethnic Group	197	34,1	34,1	34,1
		2 All Others	381	65,9	65,9	100,0
		Total	578	100,0	100,0	
3,00	Surkhet	Valid				
		1 Selected Ethnic Group	280	48,4	48,4	48,4
		2 All Others	298	51,6	51,6	100,0
		Total	578	100,0	100,0	
4,00	Banke	Valid				
		1 Selected Ethnic Group	127	22,0	22,0	22,0
		2 All Others	451	78,0	78,0	100,0
		Total	578	100,0	100,0	

For explorative purposes, however, we may treat the survey population, where each district counts the same in the final analysis. It is recommended to always use the weight_d variable if we do not split the analysis on target and non-target group.

This has implications on the results. See for example results with and without applying weights for proportion of households respectively with and without Television (C20g) in the four districts. If weights are not applied:

c20g Amenity - Television

district	Survey district		Frequency	Percent	Valid Percent	Cumulative Percent
1,00	Dhanusa	Valid				
		1 Yes	145	12,2	12,5	12,5
		2 No	1015	85,5	87,5	100,0
		Total	1160	97,7	100,0	
		Missing System	27	2,3		
		Total	1187	100,0		
2,00	Sindhupawichuk	Valid				
		1 Yes	118	21,6	22,0	22,0
		2 No	419	76,6	78,0	100,0
		Total	537	98,2	100,0	
		Missing System	10	1,8		
		Total	547	100,0		
3,00	Surkhet	Valid				
		1 Yes	53	9,0	9,0	9,0
		2 No	534	90,4	91,0	100,0
		Total	587	99,3	100,0	
		Missing System	4	,7		
		Total	591	100,0		
4,00	Banke	Valid				
		1 Yes	129	22,8	23,4	23,4
		2 No	422	74,7	76,6	100,0
		Total	551	97,5	100,0	
		Missing System	14	2,5		
		Total	565	100,0		

If applying weights:

c20g Amenity - Television

district Survey district				Frequency	Percent	Valid Percent	Cumulative Percent
1,00	Dhanusa	Valid	1 Yes	213	18,4	19,0	19,0
			2 No	910	78,7	81,0	100,0
			Total	1123	97,1	100,0	
		Missing	System	33	2,9		
		Total		1156	100,0		
2,00	Sindhupawichuk	Valid	1 Yes	137	23,7	24,3	24,3
			2 No	428	74,1	75,7	100,0
			Total	565	97,8	100,0	
		Missing	System	13	2,2		
		Total		578	100,0		
3,00	Surkhet	Valid	1 Yes	70	12,2	12,2	12,2
			2 No	506	87,6	87,8	100,0
			Total	577	99,8	100,0	
		Missing	System	1	,2		
		Total		578	100,0		
4,00	Banke	Valid	1 Yes	165	28,5	29,0	29,0
			2 No	404	69,9	71,0	100,0
			Total	569	98,4	100,0	
		Missing	System	9	1,6		
		Total		578	100,0		

Exercise: Check differences in other results when applying or not applying weights. How do you interpret the differences in results?

One can also choose to apply weights for correction of differences between analysis of:

1. Randomly selected individuals
2. All members of households

as these groups have different probabilities of being selected. However, since household size is not closely connected with key exclusion variables (tested in the survey) and application of such weights would complicate the analysis further, it was chosen not to apply such weights. Moreover, the small number of missing households made it unnecessary to apply weights for missing values.³

³ For more on the application of weights for household surveys, see for example <http://help.pop.psu.edu/help-by-statistical-method/weighting/sampling-weights-literature-review>.

4 Univariate analysis

Univariate analysis involves an examination across cases of one variable at a time. Usually we concentrate on the following three major characteristics of a single variable:

- the distribution
- the central tendency
- the dispersion

Let us go through all these characteristics for a single variable in our study:

4.1 The distribution

The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of respondents who had each value. We can for example describe the distribution of respondents in terms of their sex or their educational level. This is done by listing the number or percentage of respondents of each sex, or with different educational levels. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. With variables that can have a large number of possible values (for example income, B14), with relatively few people having each value, we group the raw scores into categories according to ranges of values (you need to know how to recode variables to do this, and if you don't, you could find it in a manual on SPSS).

One of the most common ways to describe a single variable is to make a frequency distribution. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first. For variables such as age (B4), income (B14), total working days (B16), it is not sensible to determine the frequencies for each value. Rather, the values are grouped into ranges and the frequencies determined for each range of values.

Frequency distributions can be depicted in two ways, as a table or as a graph. The table below shows an age frequency distribution with five categories of defined age ranges based on variable B4.

Frequencies

[DataSet3] H:\Nepal\methods workshop\cnas survey.sav

Statistics

broadage Broad Age Group

N	Valid	18665
	Missing	0

broadage Broad Age Group

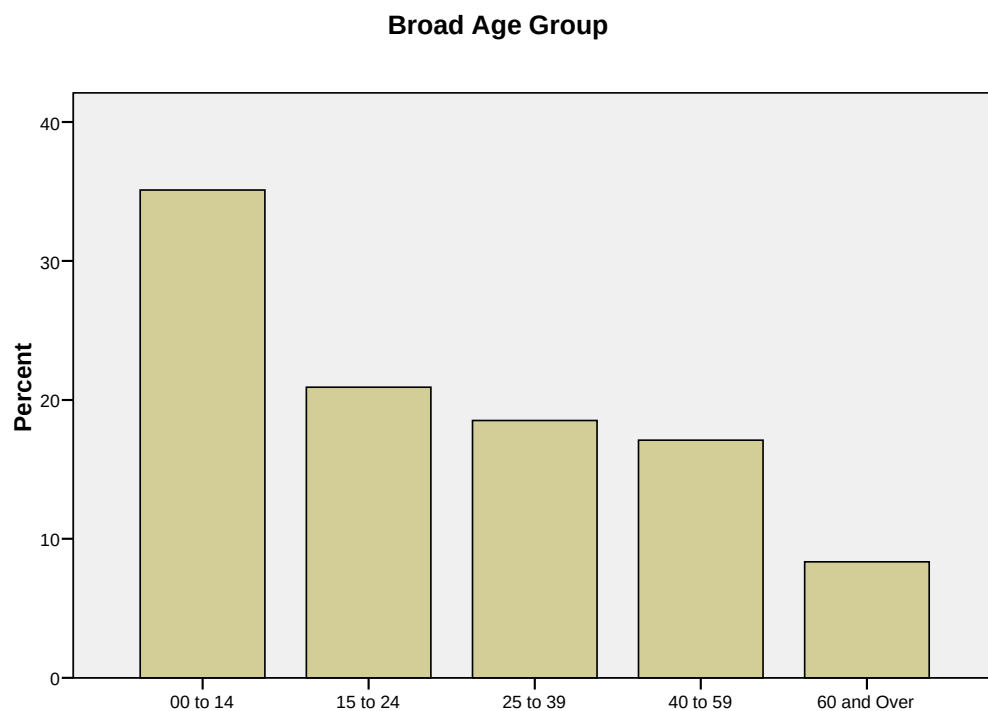
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 00 to 14	6549	35,1	35,1	35,1
	2 15 to 24	3902	20,9	20,9	56,0
	3 25 to 39	3455	18,5	18,5	74,5
	4 40 to 59	3191	17,1	17,1	91,6
	5 60 and Over	1559	8,4	8,4	100,0
	Total	18656	100,0	100,0	
Missing	0 Age Not Reported	9	,0		
Total		18665	100,0		

Note that those who have not reported their age are defined as missing value. This is done in the variable view of the data window in SPSS.

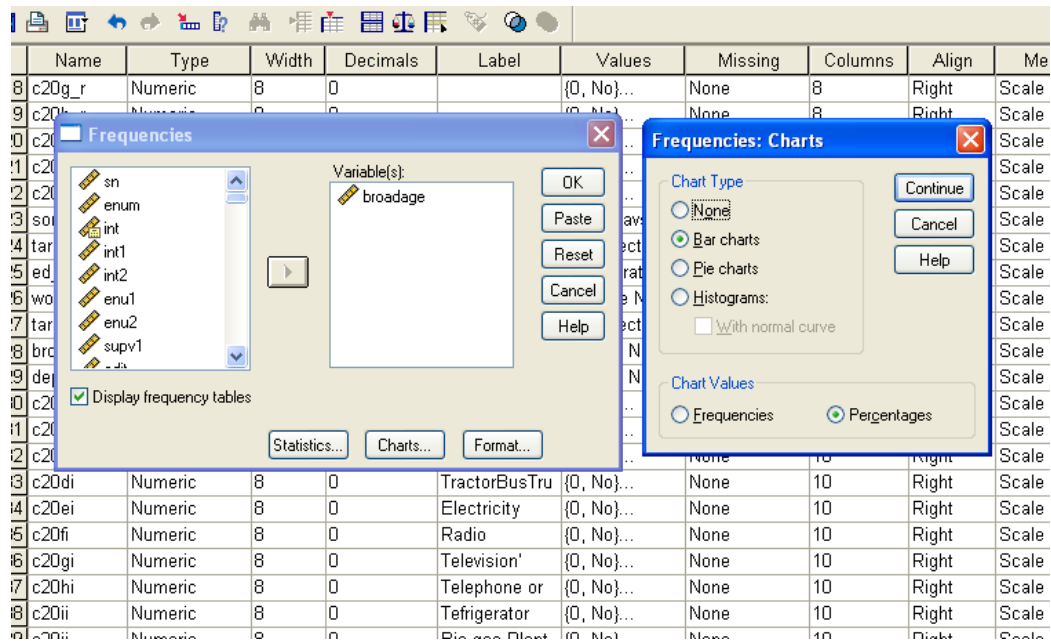
The screenshot shows the SPSS Variable View window with a list of variables. A 'Missing Values' dialog box is open, showing the 'Discrete missing values' option selected. The dialog box has fields for 'Low', 'High', and 'Discrete value'. The 'Values' column in the background shows that for the 'broadage' variable, the value '0' is defined as missing.

Name	Type	Width	Decimals	Label	Values	Missing	Column
918 c20g_r	Numeric	8	0		{0, No}...	None	8
919 c20h_r	Numeric	8	0		{0, No}...	None	8
920 c20i_r	Numeric	8	0		{0, No}...	None	8
921 c20j_r	Numeric	8	0		{0, No}...	None	8
922 c20k_r	Numeric	8	0		{0, No}...	None	8
923 sorfor2	Numeric	8	0	Survey	{1, Yadavs - D	None	8
924 target0	Numeric	8	0	on in T	{1, Selected Et	None	8
925 ed_cat	Numeric	8	0	y of Ed	{1, Illiterate}...	None	8
926 workag	Numeric	8	0	up - Pr	{0, Age Not R	None	8
927 target0	Numeric	8	0	on in T	{1, Selected Et	None	10
928 broadage	Numeric	8	0	Age Gro	{0, Age Not Re	0	10
929 depage	Numeric	8	0	Dependency b	{0, Age Not Re	None	10
930 c20ai	Numeric	8	0	Bicycle	{0, No}...	None	10
931 c20bi	Numeric	8	0	Motorcycle	{0, No}...	None	10
932 c20ci	Numeric	8	0	Car Geep	{0, No}...	None	10
933 c20di	Numeric	8	0	TractorBusTru	{0, No}...	None	10
934 c20ei	Numeric	8	0	Electricity	{0, No}...	None	10

The same frequency distribution can be illustrated in a graph as shown below. This type of graph is often referred to as a histogram or bar chart.



SPSS allows for a variety of different types of graphs to present our data. For these simple histograms, you simply click on Charts under the Frequency command and click for Bar Charts:



Distributions are usually displayed using percentages. We will come back with some additional hints on presenting the data in e.g. graphs in the final section of the paper.

EXERCISE: Use the frequency and find the

- percentage of respondents with different income levels (remember B20 = 2)
- percentage of respondents in different age ranges

4.2 Central tendency

The central tendency of a distribution is an estimate of the "centre" of a distribution of different values. There are three major types of estimates of central tendency:

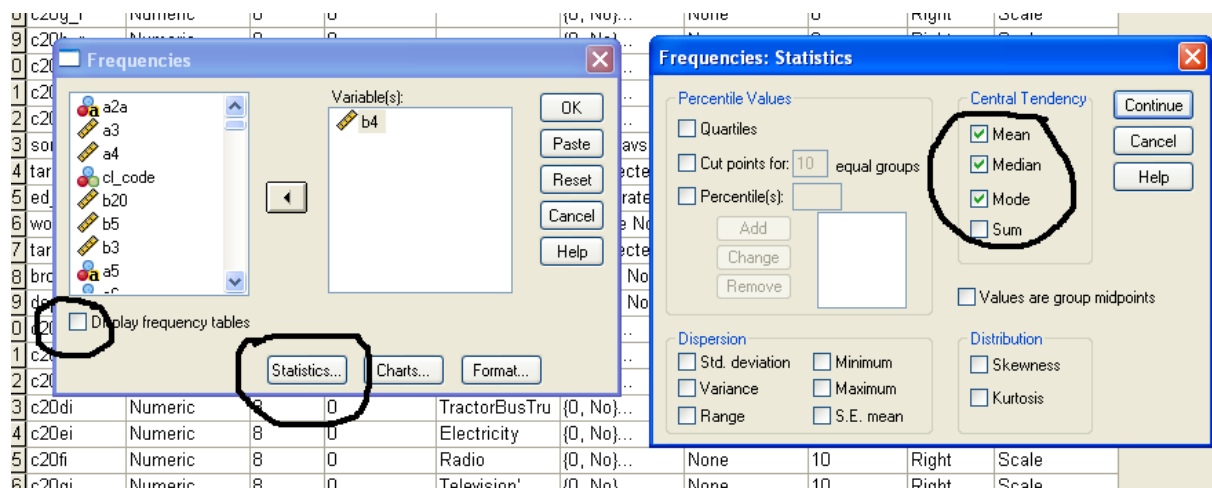
- Mean
- Median
- Mode

The **mean** (or average) is probably the most commonly used method of describing central tendency.

The **median** is the score found at the exact middle of the set of values.

The **mode** is the most frequently occurring value in the set of scores.

We can get the mean, median and mode by using the frequencies command in SPSS. The following is an illustration of how to estimate these values for the age variable (B4):



For a continuous variable (such as age) with many values, you usually don't want to display the frequency table, so make sure that the Display frequency tables is not ticked.

4.3 Dispersion

Dispersion refers to the spread of the values around the central tendency. The **Standard Deviation** is the most common, the most accurate and a very detailed estimate of dispersion. The standard deviation can be defined as:

the square root of the sum of the squared deviations from the mean divided by the number of scores minus one.

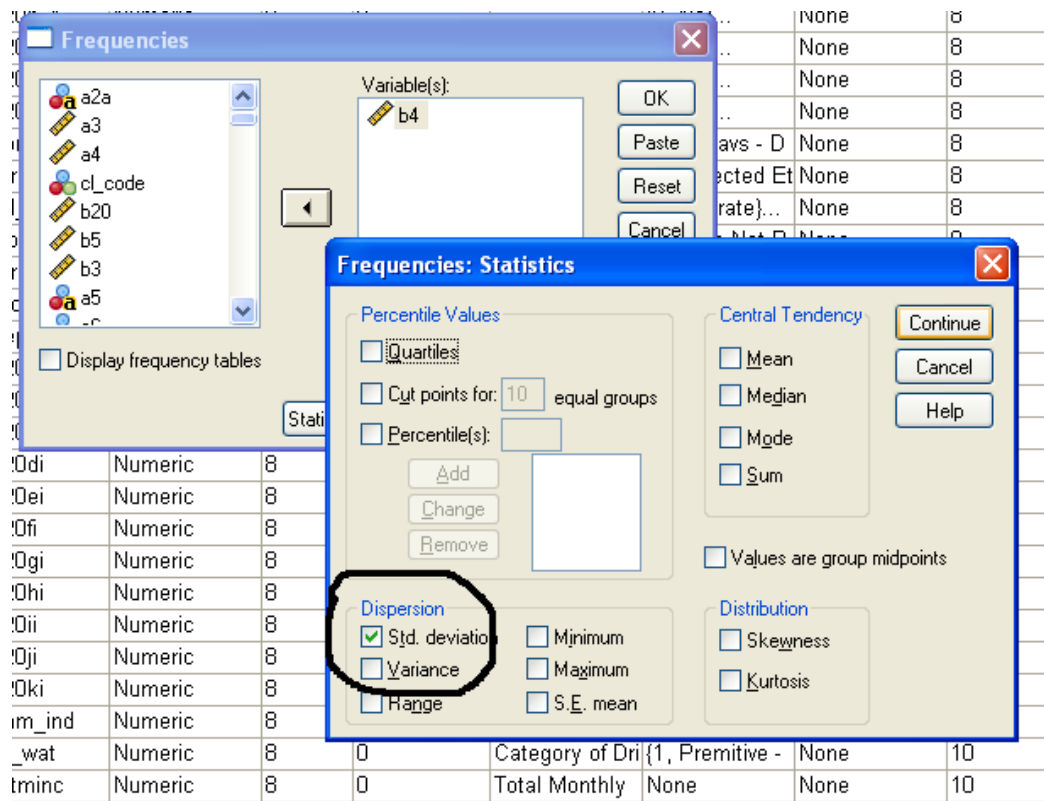
SPSS is capable of calculating the standard deviation for our variables.

The standard deviation allows us to reach some conclusions about specific scores in our distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it), then:

- approximately 68% of the scores in the sample fall within one standard deviation of the mean
- approximately 95% of the scores in the sample fall within two standard deviations of the mean
- approximately 99% of the scores in the sample fall within three standard deviations of the mean

This information enables us to compare the performance of an individual on one variable with their performance on another, even when the variables are measured on entirely different scales.

We can find the standard deviation using the frequency command:



The table below shows the mean, median, mode, minimum, maximum and standard deviation for the age variable:

Statistics

b4 Complete age

N	Valid	18665
	Missing	0
Mean		26,07
Median		21,00
Mode		10
Std. Deviation		19,689
Minimum		0
Maximum		111

Note the maximum of 111 – is it a realistic value in Nepal, or is it an outlier (error) that should be recorded as a missing value?

5 Comparing groups: Bivariate analysis

Much of what we are interested in when analysing the CNAS survey data is to compare groups of the population in terms of their risk of social exclusion for a set of indicators. Key variables for comparison are:

1. Target and non-target groups in each district
2. Districts

In addition, we can compare groups based on a large number of variables such as age, educational level, household size and composition (dependency ratio in household, male or female household head), urban/rural settlement, ethnicity, caste, religious affiliation, income levels, economic status, land ownership, and so on. We can use descriptive statistics to do so.

Inferential statistics test hypotheses about the data and may permit us to generalize beyond our data set. Examples include comparing means (averages) for a given measurement between several different groups.

The simplest form of comparing groups is to use the split-file command (remember to apply weights) and to obtain frequency, means, standard deviation, etc. for the four districts separately:

Let us first do a frequency distribution to find out if having a source of water in the house-yard is more common in certain districts than in others.

The results (after split file by district and weight by weight_d⁴) is shown in the following table:

⁴ See previous sections for how to do this.

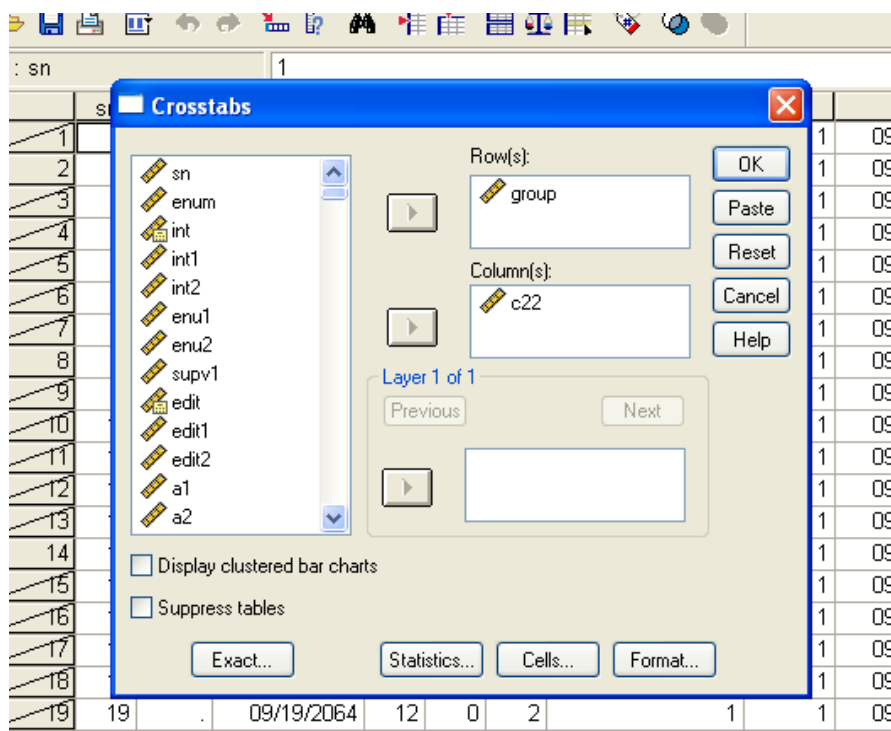
c22 Availability - Source of Water in Home-yard

district	Survey district		Frequency	Percent	Valid Percent	Cumulative Percent
1,00	Dhanusa	Valid 1 Yes	675	58,4	58,4	58,4
		2 No	481	41,6	41,6	100,0
		Total	1156	100,0	100,0	
2,00	Sindhupawichuk	Valid 1 Yes	192	33,3	33,3	33,3
		2 No	386	66,7	66,7	100,0
		Total	578	100,0	100,0	
3,00	Surkhet	Valid 1 Yes	122	21,0	21,0	21,0
		2 No	456	79,0	79,0	100,0
		Total	578	100,0	100,0	
4,00	Banke	Valid 1 Yes	466	80,6	80,6	80,6
		2 No	112	19,4	19,4	100,0
		Total	578	100,0	100,0	

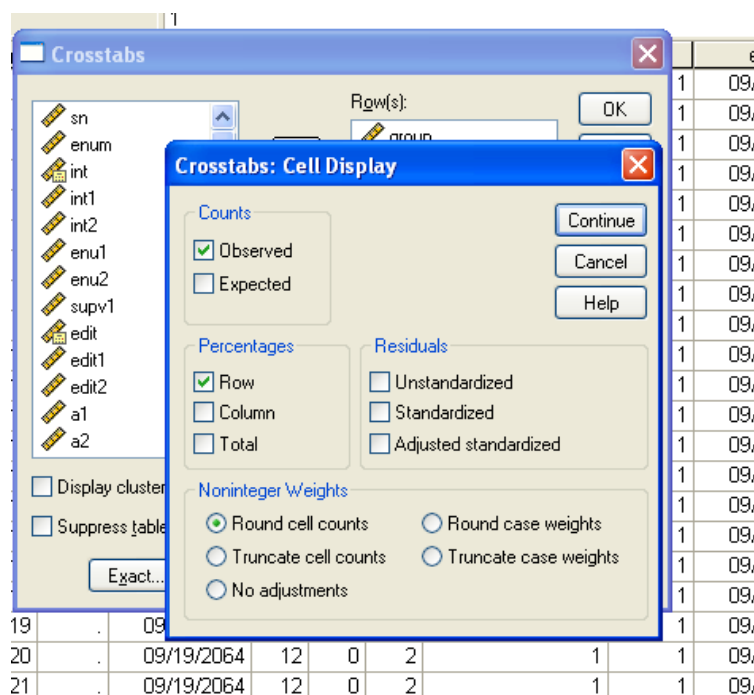
It shows distinct district-wise differences.

Let us now proceed to see if our target groups are more or less likely to have source of water than the rest of the population. We can use the cross-tabs command to do this:

In the row-field we enter the *group* variable, in the column box we enter C22.



We click on Cells, and then click on observed counts and Row percentages to get percentages as well as the observed cases:



We can also click on statistics – but will come back to this later.

The results we get are the following:

group * c22 Availability - Source of Water in Home-yard Crosstabulation

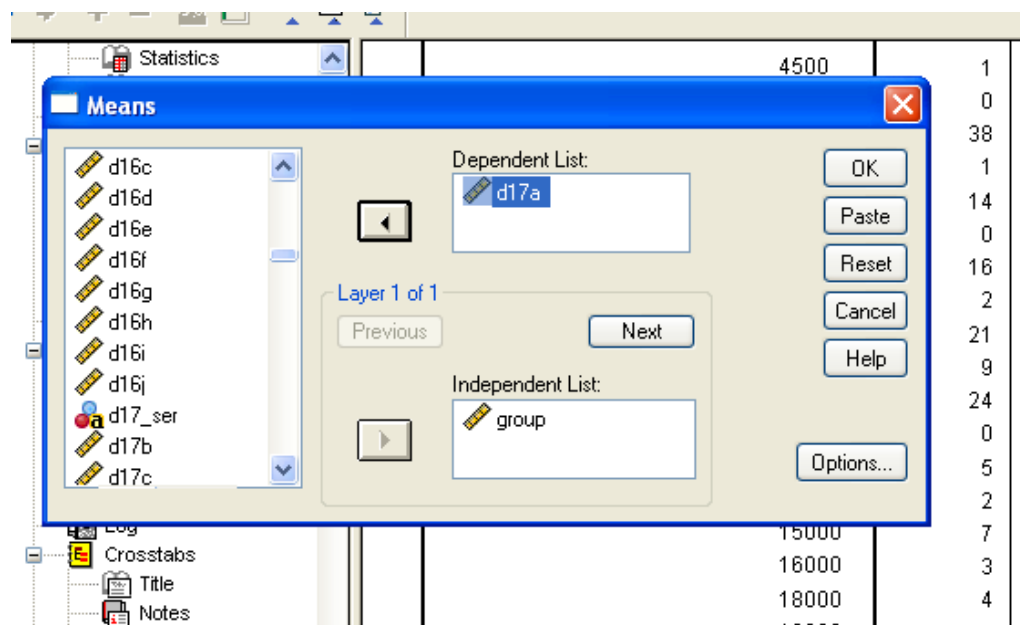
					c22 Availability - Source of Water in Home-yard		Total
district Survey district					1 Yes	2 No	
1,00 Dhanusa	group	1,00 Yadavs. Dhanusa	Count		123	81	204
			% within group		60,3%	39,7%	100,0%
		2,00 Tarai Dalits. Dhanusa	Count		29	70	99
			% within group		29,3%	70,7%	100,0%
		3,00 Others. Dhanusa	Count		524	330	854
			% within group		61,4%	38,6%	100,0%
	Total		Count		676	481	1157
			% within group		58,4%	41,6%	100,0%
2,00 Sindhupawlichuk	group	4,00 Tamangs. Sindhupalchowk	Count		56	130	186
			% within group		30,1%	69,9%	100,0%
		5,00 Others. Sindhupalchowk	Count		136	256	392
			% within group		34,7%	65,3%	100,0%
	Total		Count		192	386	578
			% within group		33,2%	66,8%	100,0%
3,00 Surkhet	group	6,00 Hill Dalits. Surkhet	Count		22	99	121
			% within group		18,2%	81,8%	100,0%
		7,00 Others. Surkhet	Count		99	358	457
			% within group		21,7%	78,3%	100,0%
	Total		Count		121	457	578
			% within group		20,9%	79,1%	100,0%
4,00 Banke	group	8,00 Muslims. Banke	Count		104	18	122
			% within group		85,2%	14,8%	100,0%
		9,00 Others. Banke	Count		362	94	456
			% within group		79,4%	20,6%	100,0%
	Total		Count		466	112	578
			% within group		80,6%	19,4%	100,0%

We can see rather large differences between groups. The highest share of those with source of water in the home-yard are found among Muslims and Others in Banke, then Yadavs and Others in Dhanusa. The lowest percentage is found among respondents in Surkhet, regardless of their group belonging.

Exercise: Find group differences between target and non-target groups in each district in terms of household ownership of land (C1).

Let us say that we are interested in finding the mean amount of Nepali rupies spent on health care in households during the past year by district and target/non-target group.

In the Data window, go to the Analyze menu, select Compare Means and enter as follows:



You then get the following table, indicating highest average health care expenses for Yadav households in Dhanusa, followed by Others in Sindhupawlchuk. The lowest are found among Tamangs in Sindhupawlchuk, Hill and Tarai Dalits in Surkhet. It is worth noting that Muslims in Banke have no lower average than other groups.

Report

d17a Health Care

district	Survey district	group	Mean	N	Std. Deviation
1,00	Dhanusa	1,00 Yadavs. Dhanusa	13398,14	203	26007,110
		2,00 Tarai Dalits. Dhanusa	5645,34	98	13385,475
		3,00 Others. Dhanusa	7752,20	854	13128,832
		Total	8566,81	1156	16319,144
2,00	Sindhupawlchuk	4,00 Tamangs. Sindhupalchowk	5027,09	186	12495,352
		5,00 Others. Sindhupalchowk	8659,13	392	21264,489
		Total	7489,61	578	18955,244
3,00	Surkhet	6,00 Hill Dalits. Surkhet	5491,75	121	13221,752
		7,00 Others. Surkhet	8500,25	457	26371,171
		Total	7871,47	578	24241,196
4,00	Banke	8,00 Muslims. Banke	8404,75	122	20394,401
		9,00 Others. Banke	6124,24	456	8691,282
		Total	6605,43	578	12150,403

5.1 Bivariate measures of association and significance tests

So far we have given descriptive bivariate statistics. But – as mentioned above – in our research papers we often wish to make inferences from the sampled population to the population as a whole. In the CNAS survey we can do this to some extent, but we should also do so with great caution due to:

1. We have drawn a sample only from four districts of Nepal.
2. The sample design is complex, while significance tests conducted in SPSS assume simple random sampling.⁵
3. Some groups are overrepresented in the survey. This is compensated by weights, but affects significance tests.
4. The sample is drawn from villages with a certain proportion of both target and non-target ethnic groups, while mono-ethnic environments were not included.

These conditions should not, however, restrict us from conducting significance tests and measure the strength of association between variables. Even if our results are not completely accurate, they nevertheless give a good indication of the correlation between variables and to what extent we are able to draw conclusions from our findings. A precaution would be to require a stronger association and require a lower significance level than we would normally do if we had drawn a completely random sample. For example, while confidence intervals are usually set to 95% - and significance tests are based upon 5% significance levels, these could be increased to 99% and 1% respectively to compensate for the described imprecision.

⁵ There is software available, also in SPSS, which handles complex sample designs, but such software is yet not available to researchers in the project.

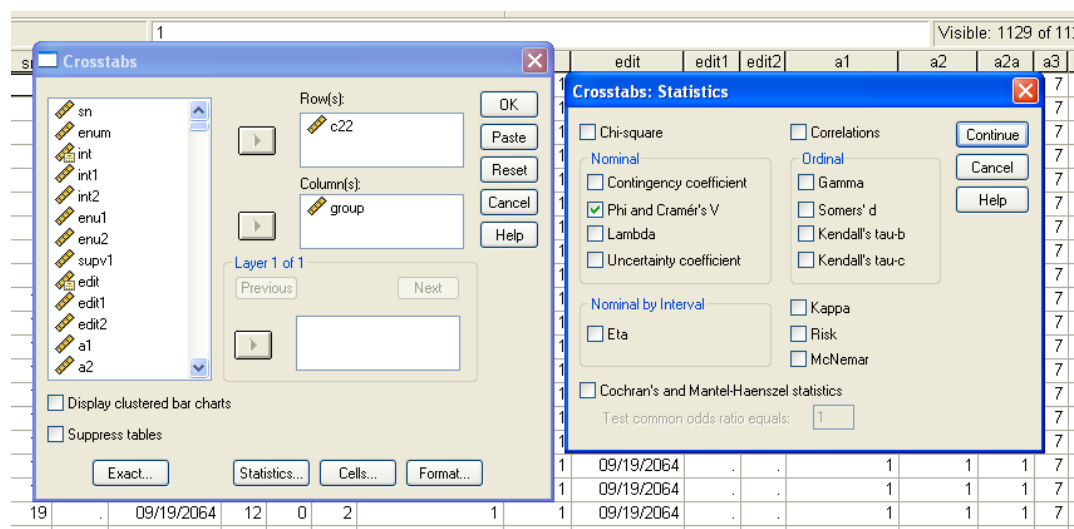
We should also be open about the limitations to readers of our analysis, and for example not argue that we can draw conclusions about the whole country of Nepal.

Let us now go back to the two examples above and look at measures of association between the variables.

Which measures that are appropriate to use depends on the measurement level (nominal, ordinal or scale (interval/ratio)).

A research question could for example be formulated as follows: “Is source of water in the house-yard associated with group belonging (target vs non-target groups)?” Our preliminary finding showed rather large differences between groups in Dhanusa, but not so big differences between groups in Sindhupawlhuk, Surkhet and Banke. It seems district differences are larger than group differences in the districts, with an exception for Dhanusa.

We want to test the null hypothesis that there is no difference between groups. For this analysis we have variables at the nominal level, and Phi / Cramer's V are appropriate. We select Crosstabs again, and click on the box for Statistics, and then tick the box for Phi and Cramer's V.



The result is shown below:

Symmetric Measures

		Value	Approx. Sig.
Nominal by	Phi	,436	,000
Nominal	Cramer's V	,436	,000
N of Valid Cases		2891	

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

This shows statistically significant associations between group belonging and likelihood of having a source of water in the house-yard. However, if we do district-wise analysis (which we should do according to our sample design), we get the following result:

Symmetric Measures

district Survey district			Value	Approx. Sig.
1,00 Dhanusa	Nominal by	Phi	,181	,000
	Nominal	Cramer's V	,181	,000
	N of Valid Cases		1157	
2,00 Sindhupawichuk	Nominal by	Phi	-,045	,274
	Nominal	Cramer's V	,045	,274
	N of Valid Cases		578	
3,00 Surkhet	Nominal by	Phi	-,035	,403
	Nominal	Cramer's V	,035	,403
	N of Valid Cases		578	
4,00 Banke	Nominal by	Phi	,060	,146
	Nominal	Cramer's V	,060	,146
	N of Valid Cases		578	

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

Only in Dhanusa are there statistically significant differences between target and non-target groups. It seems that differences between districts are more important in explaining variation between groups than differences between target and non-target groups in districts. This is strengthened by the following table with association between district and C22:

Symmetric Measures

		Value	Approx. Sig.
Nominal by	Phi	,419	,000
Nominal	Cramer's V	,419	,000
N of Valid Cases		2890	

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

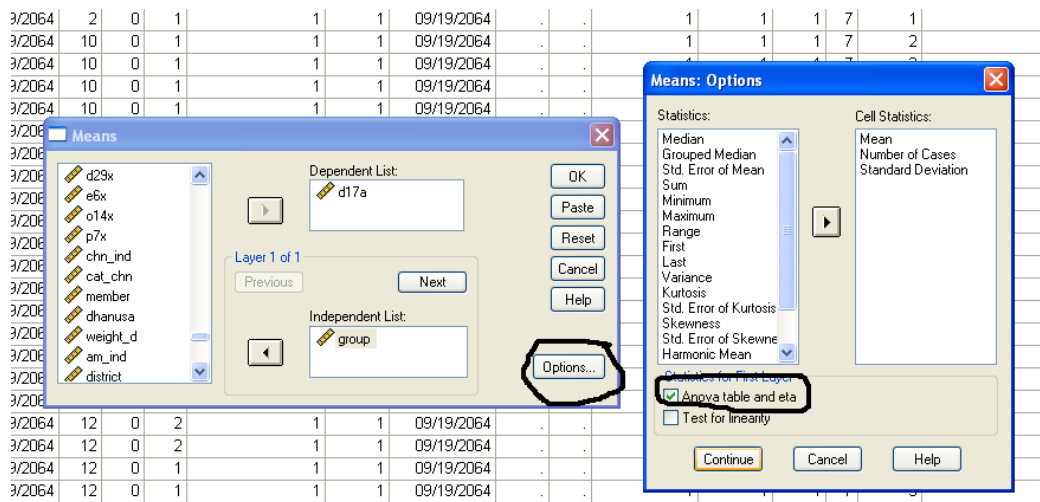
The association (measured by *Phi* and *Cramer's V*) are almost equally large between district and C22 as between group and C22.

Phi and *Cramer's V* are appropriate to use when we deal with two nominal variables (C22 can be considered both a nominal and an ordinal variable).

When we come to nominal by scale (as is the case with group/district (nominal) and health care expenses (scale)) we use other measures of association.

Our research question is to find out whether household expenses to health care (D17a) are associated with group affiliation and/or district. Eta is the appropriate measure for this.

Go to the *Compare Means* under the *Analyze* scroll-down menu. Click *Options...* and then tick the *Anova table and eta* in the window that comes up, then *Continue* and *OK*.



The results give an Eta squared of 0.11, which – as shown in the ANOVA Table – is a statistically significant result. The derived output indicates a high likelihood that the association between the group belonging and health care expenses will be present in the population. Thus, it is highly likely that this association is found not only in our sample but exists in the real world in our four districts combined.

8,00 Muslims. Banke	8404,75	122	20394,401
9,00 Others. Banke	6124,24	456	8691,282
Total	7820,03	2890	18044,996

ANOVA Table

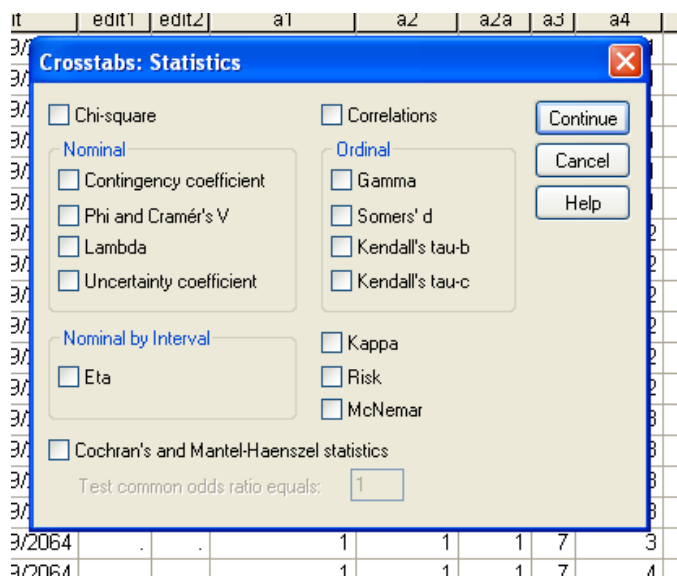
	Sum of Squares	df	Mean Square	F	Sig.
d17a Health Care * group	1,1E+010	8	1343312066	4,160	,000
Within Groups	9,3E+011	2880	322908020,1		
Total	9,4E+011	2888			

Measures of Association

	Eta	Eta Squared
d17a Health Care * group	,107	,011

Exercises: Are there statistically significant district-level differences? Are differences between groups statistically significant in all districts (split file)?

You now have the tools to conduct bivariate analysis for different types of variables. The box in the statistics window shows what types of measurements are appropriate for different types of variables.



However, consult statistics handbooks to be sure that you apply the correct measures and for how to interpret the results. One general guide is the following⁶:

Bivariate Associations: Level of Measurement vs Measure of Association			
	Nominal	Ordinal	Interval/Ratio
Nominal	Contingency C Cramers V Lambda Phi Yules Q	Contingency C Cramers V Lambda Phi Yules Q	Eta Square
Ordinal	Contingency C Cramers V Lambda Phi Yules Q	Gamma Somers d Tau a Tau b Tau c	A range of other measures are used here but we will not discuss them
Interval/Ratio	Eta Square	A range of other measures are used here but we will not discuss them	Pearson's r (i.e. the correlation coefficient)

⁶ From http://salises.mona.uwi.edu/sem1_08_09/SALI6012/Data_Analysis/Data%20Analysis.pdf.

6 Creating additive indexes

A concept is usually much richer than any single measure of it. Therefore both reliability and validity may be enhanced by developing a number of measures of the same underlying concept and then combining them into a scale or index.

An index can be created simply by adding the values of the individual measures that make it up. For example, in the CNAS survey, there is a question (G1) asking about access to facilities. Any person could either answer yes or no of each of the facilities. By adding up the number of positive answers, one would presumably get an index of access to facilities, which is better than any single item.

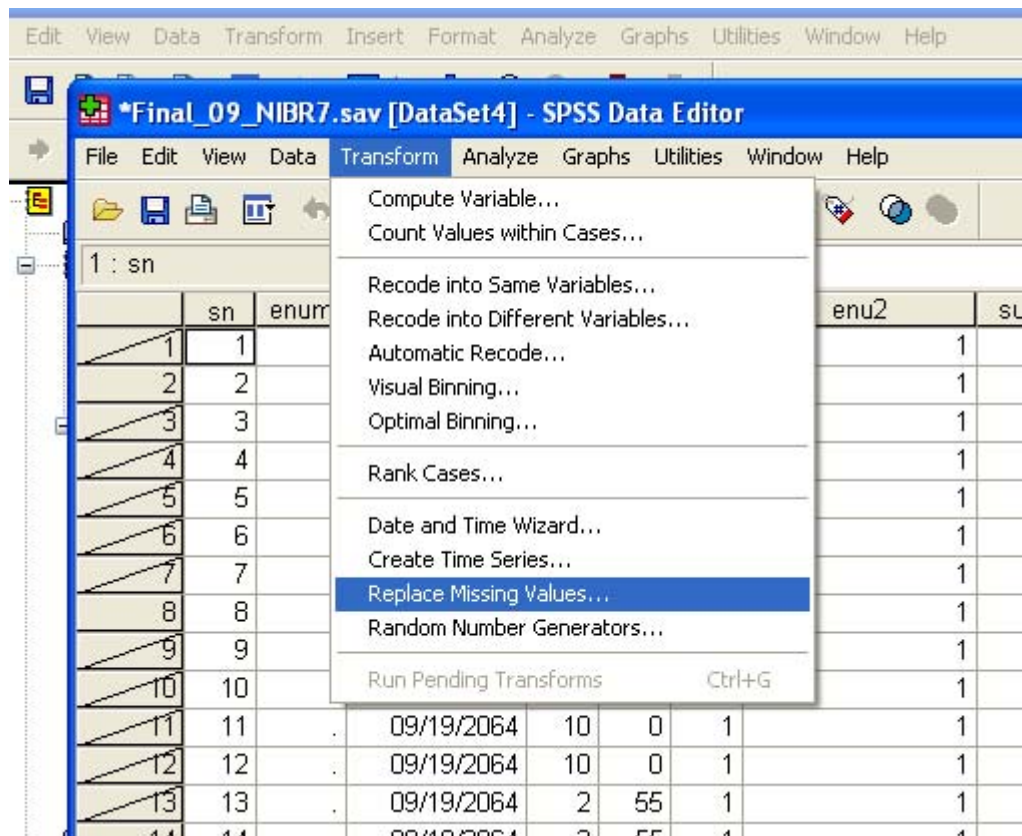
How do we do this in practice?

First we take a look at the distribution of responses. Remember that Select cases (B20 = 2) should be selected. The responses are 1 'yes', 2 'no', 8 'do not know' and missing. First we rearrange (recode) so that 'no' = 0 and don't know is defined as a

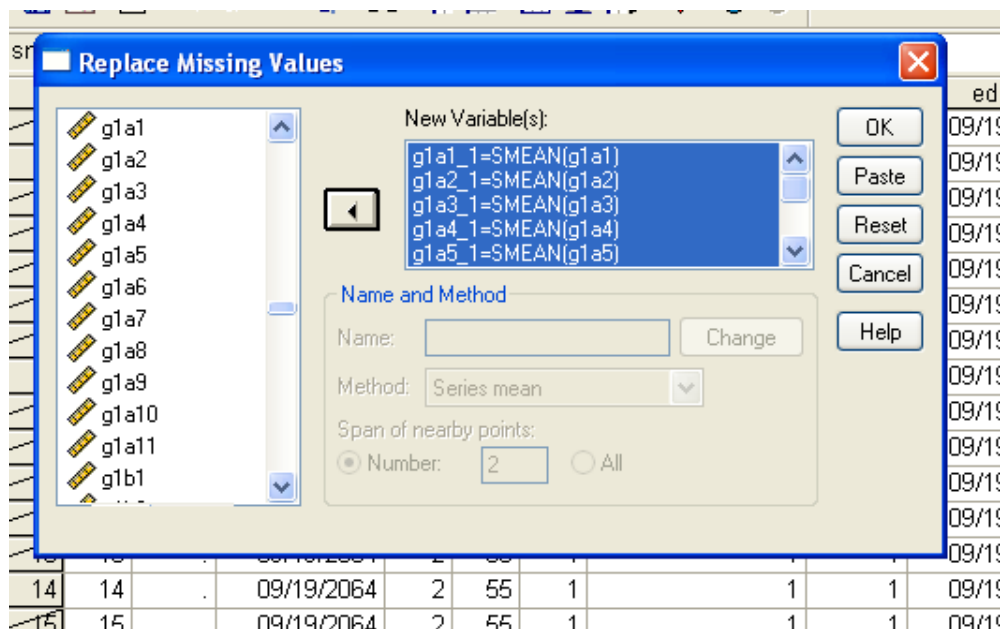
The syntax for doing this is:

```
RECODE
  g1a1 g1a2 g1a3 g1a4 g1a5 g1a6 g1a7 g1a8 g1a9 g1a10 g1a11 (2=0) .
EXECUTE .
VALUE LABELS g1a1 g1a2 g1a3 g1a4 g1a5 g1a6 g1a7 g1a8 g1a9 g1a10 g1a11 1 'Yes'
0 'No' 8 'Do not know'.
MISSING VALUES g1a1 g1a2 g1a3 g1a4 g1a5 g1a6 g1a7 g1a8 g1a9 g1a10 g1a11 (8).
```

We cannot assume that all the missing values don't have access. We have two options, either exclude them from the analysis (that means, that if a respondent for some reason has a missing value for only one of the 11 items, he or she will be excluded from this index), or create new variables, where the missing values and the don't know are ascribed the average number of all the other responses. In the following example, we have ascribed the average value to missing cases (so that they will be included in other analyses).



Select the variables that you wish to use (G1a1 to G1a11) and click OK



You make the index based on these new variables.

An additive index can be created by simply adding up all the values.

```
COMPUTE amen_ind = g1a1_1 + g1a2_1 + g1a3_1 + g1a4_1 + g1a5_1 + g1a6_1 +
g1a7_1 + g1a8_1 + g1a9_1 + g1a10_1 + g1a11_1.
```

We have now created an index of access to amenities with a potential score from 0 (no amenities) to 11 (all amenities). Let us look at the central tendency and dispersion of the index:

Statistics

amen_ind		
N	Valid	2890
	Missing	0
Mean		5,6632
Median		5,5277
Mode		4,00
Std. Deviation		2,64331
Minimum		,00
Maximum		11,00

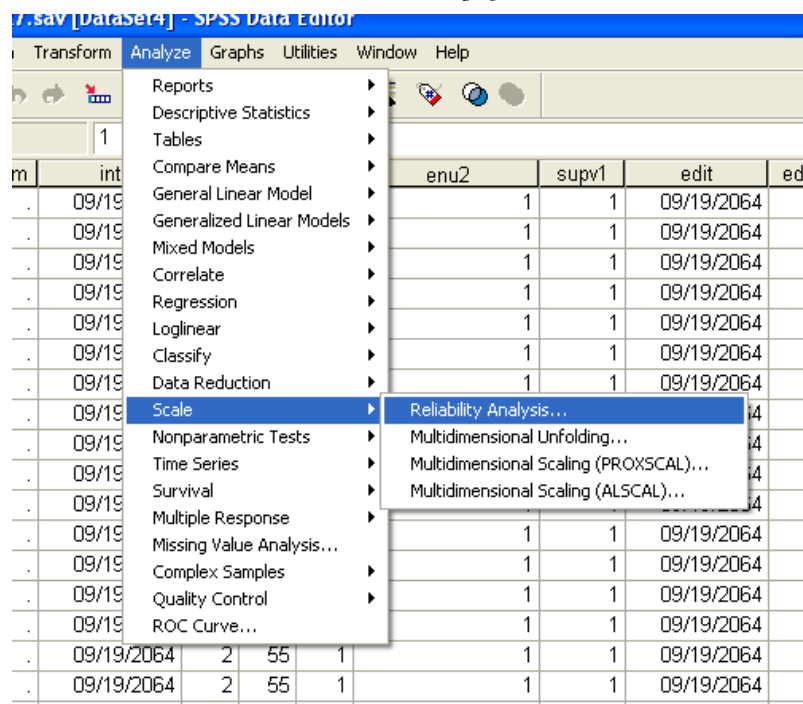
We see that the average (mean) score on the index is 5.7. Some households have access to no, while some households have access to all 11 amenities.

However, to what extent do all of the items included in the amenities index really measure the same concept? One common way to test this is to make the generally reasonable assumption that the composite index is more valid and reliable than any one of the items that make it up. We can correlate each individual item in the index with the score on the composite index. A low correlation would indicate that a particular item is not closely related to the index. That item could then be dropped, and the index recalculated.

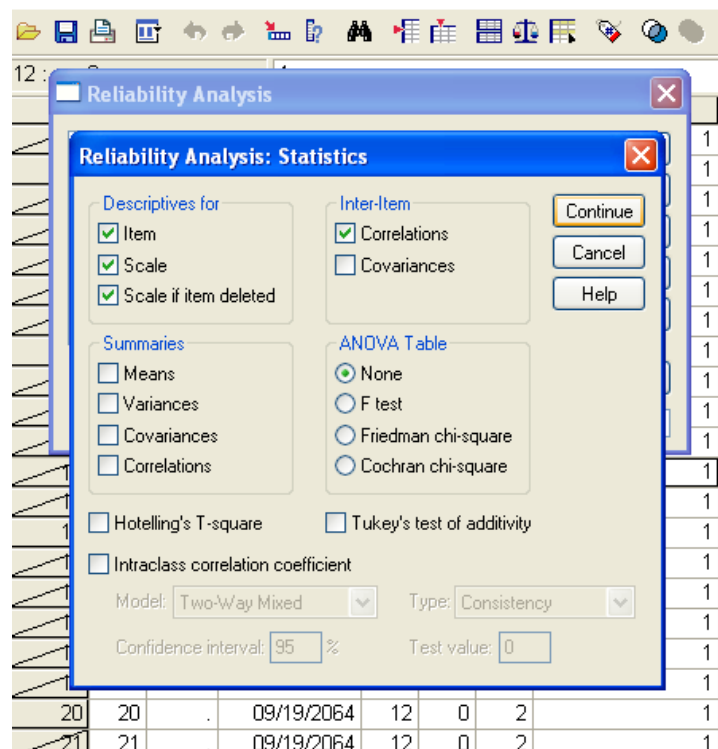
We usually also perform *reliability analysis* for the index as a whole. A commonly used measure of an index's reliability is the *Cronbach's Alpha* (α). This measure is calculated from the number of items making up the index and the average correlation among those items. The higher the value of Alpha, the more reliable the index. The value of Alpha generally ranges from zero to one. However, a negative value is technically possible. A score of at least .70 is generally considered acceptable for creating an index.

The reliability analysis can be performed in SPSS in the following way:

1. In the data window, choose *Analyze*, then *Scale*, and select *Reliability Analysis*



2. Select the 11 (new) variables in the potential index and tick the boxes as shown below and click Continue, and in the next Window OK:



The first result shows a Chronbach's Alpha of 0.78. It is above the requirement of 0.70.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,784	,774	11

However, are all items to be included in the index? Let's go to the Item-Total Statistics box:

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
g1a1_1 SMEAN(g1a1)	4,762	6,734	,107	,054	,795
g1a2_1 SMEAN(g1a2)	5,278	5,579	,520	,359	,758
g1a3_1 SMEAN(g1a3)	4,863	6,005	,424	,213	,769
g1a4_1 SMEAN(g1a4)	5,271	5,499	,550	,352	,754
g1a5_1 SMEAN(g1a5)	5,046	5,547	,530	,382	,756
g1a6_1 SMEAN(g1a6)	5,537	6,247	,382	,209	,774
g1a7_1 SMEAN(g1a7)	4,814	6,318	,303	,160	,781
g1a8_1 SMEAN(g1a8)	5,145	5,419	,575	,492	,750
g1a9_1 SMEAN(g1a9)	5,264	5,395	,603	,537	,747
g1a10_1 SMEAN(g1a10)	5,516	6,055	,474	,327	,765
g1a11_1 SMEAN(g1a11)	5,136	6,091	,288	,129	,786

One can see from the result that by removing two of the items, one would get a Chronbach's Alpha that is higher than 0.784. In order to get an index that to the largest possible extent measure one concept (access to amenities), we would consider removing g1a1_1 and g1a11_1 (drinking water and electricity) from the index. Conceptually, this makes sense, as drinking water and electricity are normally not facilities that are associated with other types of services that are listed in the index. Instead of the index above, we should therefore rather have made an index including only the other items in the list. Since it is an indicator of access to services, we change the name:

COMPUTE serv_ind = g1a2_1 + g1a3_1 + g1a4_1 + g1a5_1 + g1a6_1 + g1a7_1 + g1a8_1 + g1a9_1 + g1a10_1.

However, testing the new scale in a reliability analysis, gives a Chronbach's Alpha of 0.796 and shows that the new index would be improved by removing primary school as well.

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
g1a2_1 SMEAN(g1a2)	3,849	4,540	,544	,354	,775
g1a3_1 SMEAN(g1a3)	3,435	5,010	,401	,190	,793
g1a4_1 SMEAN(g1a4)	3,843	4,543	,533	,328	,776
g1a5_1 SMEAN(g1a5)	3,618	4,541	,538	,380	,776
g1a6_1 SMEAN(g1a6)	4,108	5,200	,380	,206	,795
g1a7_1 SMEAN(g1a7)	3,386	5,247	,311	,159	,802
g1a8_1 SMEAN(g1a8)	3,716	4,408	,591	,487	,767
g1a9_1 SMEAN(g1a9)	3,836	4,378	,626	,536	,762
g1a10_1 SMEAN(g1a10)	4,088	4,992	,494	,324	,783

One should do this exercise until one reaches the best possible index. Finally we arrive at an index with only 8 items, but with a very high internal correlation between all the items and a very high Chronbach's Alpha.

Exercise: Compute the index as shown above and find the average score on the index for target and non-target groups in each of the four districts.

Exercise: Create an additive index for ownership of household consumer goods (C20). Find the minimum, maximum and average score for target and non-target groups in each of the four districts.

7 Multivariate analysis

In this section we will go through two types of multivariate analysis (i.e. analyses where we have one dependent and more than one independent variables): Multiple and logistic regression. There are a number of other multivariate analysis techniques, but we have selected two very commonly used techniques for different types of dependent variables and suggest that you master these two ones before you proceed to more advanced techniques.

7.1 Multiple linear regression

The aim of regression analysis is to estimate the effect or impact of a given independent variable on variation in the dependent variable. In the case of multiple regression, we control for all the other independent variables in the model.

We have already made an index for accessibility of services in the community. We would like to see to what extent this level is affected by district, group affiliation, rural/urban settlement, household poverty and experienced improvements in facility level.

We use multiple linear regression to calculate how much the dependent variable (service level) changes when other variables (independent) change.

Here we assume some previous knowledge of multiple linear regression. If you are not familiar with regression analysis, you should first consult a statistics textbook. Our aim is to show you how to perform such analysis in SPSS for Windows with the CNAS data set.

The **dependent** variable is serv_ind (service index).

Independent variables are:

A2 (high: urban; low: rural)

Group: caste (all caste groups), dalit, janjati and muslim

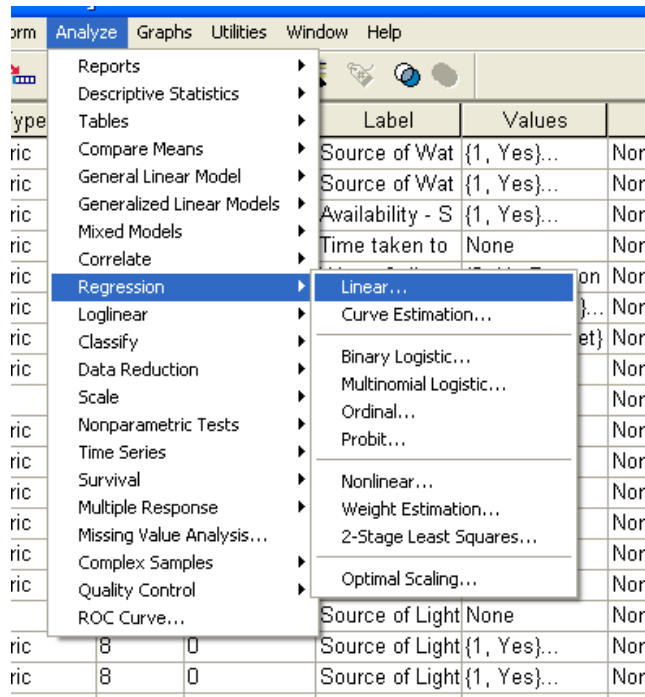
District: d_dhan, d_sindhu, d_surkh, d_banke

Poverty: low_income: among the 20% households with lowest income

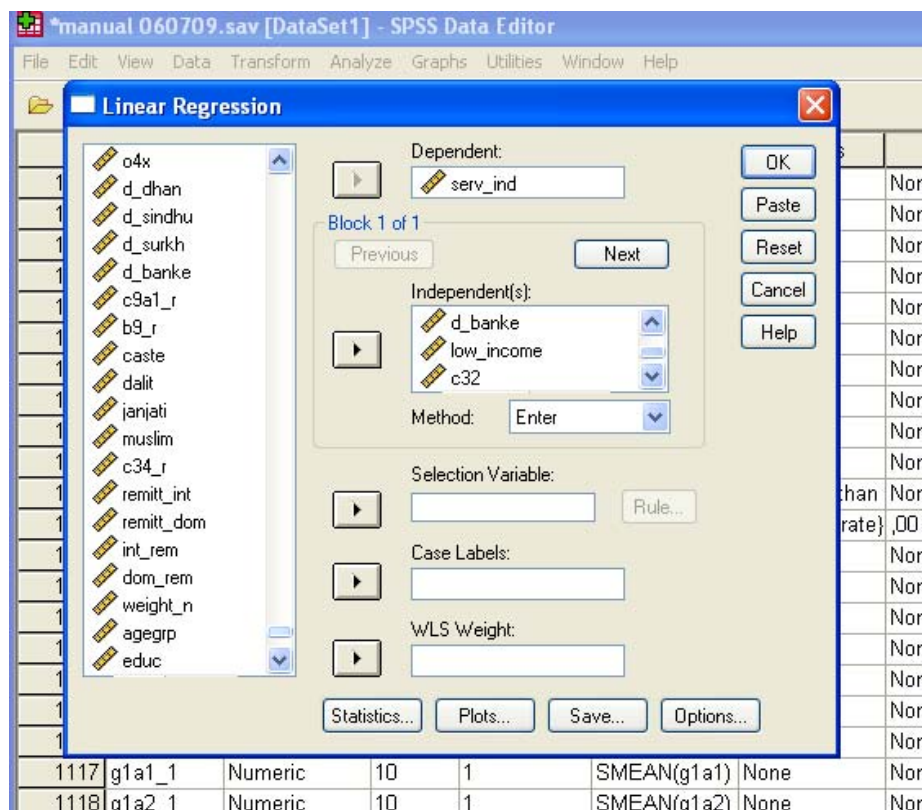
C32: experienced improvement (low: much improvement).

Note that groups and districts are converted into dichotomous (dummy) variables.

First, in the data file choose *Analyze* in the scroll-down menu, then select *Regression* and *Linear*



In the window that appears, select the dependent variable (`serv_ind`) and the independent. You may wish to run optional analyses, such as checking for collinearity, histograms, etc., but we will not do so here.



For different types of methods (step-wise, forward, backward, etc.), consult statistics handbooks. Here we use the default Enter method (all independent variables are entered simultaneously into the model).

Let us first look at the model summary:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,335(a)	,112	,109	2,19886

a Predictors: (Constant), c32 Household Facilities Compared - Intergenerational, d_surkh, janjati, a2 VDC/Municipality, low_income Among the lowest 20% per capita household income, muslim, d_banke, dalit, d_sindhu

In a multiple linear regression model, adjusted R square measures the proportion of the variation in the dependent variable accounted for by the explanatory variables. Unlike R square, adjusted R square allows for the degrees of freedom associated with the sums of the squares. Adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square (they are very similar in our case, however). Thus, approximately 11 per cent of the variation in terms of availability of services is explained by the independent variables in the model.

The anova table tests the acceptability of the model from a statistical perspective.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1759,612	9	195,512	40,437	,000(a)
	Residual	13924,738	2880	4,835		
	Total	15684,351	2889			

a Predictors: (Constant), c32 Household Facilities Compared - Intergenerational, d_surkh, janjati, a2 VDC/Municipality, low_income Among the lowest 20% per capita household income, muslim, d_banke, dalit, d_sindhu

b Dependent Variable: serv_ind

The Regression row displays information about the variation accounted for by our model. The Residual row displays information about the variation that is not accounted for by our model. The regression and residual sums of squares are of different sizes and confirm that about 11 per cent of the variation in amenities level is explained by the model.

The significance value of the F statistic is less than 0.05 (or 0.01 which is the significance level we have set due to the sampling imperfections explained in a previous section), which means that the variation explained by the model is not due to chance.

Let us proceed to look at the coefficient table:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3,332	,248		13,419	,000		
	a2 VDC/Municipality	1,346	,158	,154	8,508	,000	,942	1,061
	dalit	-,031	,139	-,005	-,224	,823	,760	1,315
	janjati	,095	,121	,015	,785	,432	,834	1,199
	muslim	-,227	,194	-,024	-1,173	,241	,768	1,302
	d_sindhu	,621	,121	,107	5,111	,000	,709	1,411
	d_surkh	,635	,115	,109	5,535	,000	,795	1,258
	d_banke	-,762	,119	-,131	-6,383	,000	,733	1,363
	low_income Among the lowest 20% per capita household income	-,285	,116	-,049	-2,454	,014	,777	1,287
	c32 Household Facilities Compared - Intergenerational	-,591	,064	-,167	-9,248	,000	,942	1,062

a. Dependent Variable: serv_ind

Standardized coefficients or beta coefficients are the estimates resulting from an analysis performed on variables that have been standardized so that they have variances of 1. We want to answer the question of which of the independent variables have a greater effect on the dependent variable, but know that the variables are measured in different units of measurement. From the table we can see that the Beta coefficients are highest for C32 (perceived improvements in household facilities) and A2 (urban/rural type of settlement). To determine the relative importance of the significant predictors, we should therefore rather look at the standardized than the unstandardized coefficients. Even though C32 has a smaller

coefficient than `d_sindhu` and `d_banke`, `C32` contributes more to the model because it has a larger absolute standardized coefficient.

The analysis shows that the group belonging of respondents is not a statistically significant variable in explaining different levels of availability of services in the community when other variables in the model are controlled for. This makes sense, since all people in the village, regardless of their caste, ethnicity or religion, will have services available (another matter is the extent to which they are able to use them).

Statistically significant variables, however, are urban/rural residence (people in urban areas have significantly better access) and households facilities compared with the past (those who have experienced improvements have better availability of services). Both of these findings are plausible. More interestingly, however, is the impact of district. Compared to people in Dhanusa (control group), people in Sindhupawlchuk and Surkhet have on average more services available, while people in Banke have fewer – and the results are statistically significant. Finally, people with low income tend to report lower availability of services, but the significance level is on the margin (we have defined it as 0.01, and in this case the relationship is not statistically significant).

When the **tolerances** are close to 0, there is high multicollinearity and the standard error of the regression coefficients will be inflated. A **variance inflation factor** greater than 2 is usually considered problematic, and the highest VIF in the table is 1.411. Thus, in this model we do not seem to have a problem of multicollinearity.

7.2 Logistic regression

While linear regression is useful for dependent variables at interval or ratio (scale) level, binary logistic regression is most useful when you want to model the event probability for a categorical response variable with two outcomes; typically yes or no, have or have not, etc.⁷

For example:

We would like to know what factors that explain why some people feel they have not equal opportunities as other people in their community to have access to employment in government jobs.

Our dependent variable is civil society membership (1 = not equal opportunity, 0 = equal opportunity).

First we compute a new variable which we call `job_opp` (Job opportunity), for example using this syntax:

⁷ For a more thorough introduction to logistic regression analysis, you should consult a statistics handbook.

```

recode d7 (2 = 1) (1 = 0) (else = copy) into job_opp.
missing values job_opp (3 thru high).
variable labels job_opp "Perceived employment opportunity in government".
val lab job_opp 1 'Less opportunity' 0 'Equal opportunity'.
format job_opp (F2.0).
freq job_opp.

```

The results show that only 4 in 10 of the respondents believe they have equal job opportunity.

job_opp Perceived employment opportunity in government

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Equal opportunity	980	33,9	39,9	39,9
	1 Less opportunity	1475	51,0	60,1	100,0
	Total	2454	84,9	100,0	
Missing	8	390	13,5		
	9	9	,3		
	System	37	1,3		
	Total	436	15,1		
Total		2890	100,0		

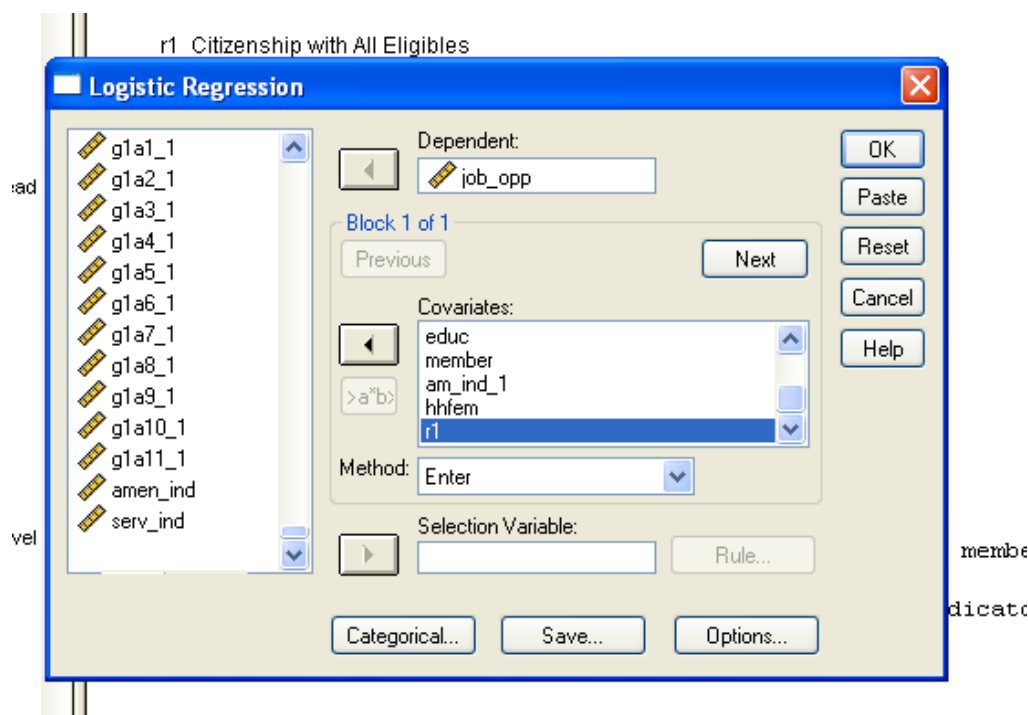
Then we think of which independent variables to include in the model. Our selection of independent variables should be guided by some assumptions about possible relationships.

For an exploratory model (which can all the time be refined), we include the following variables:

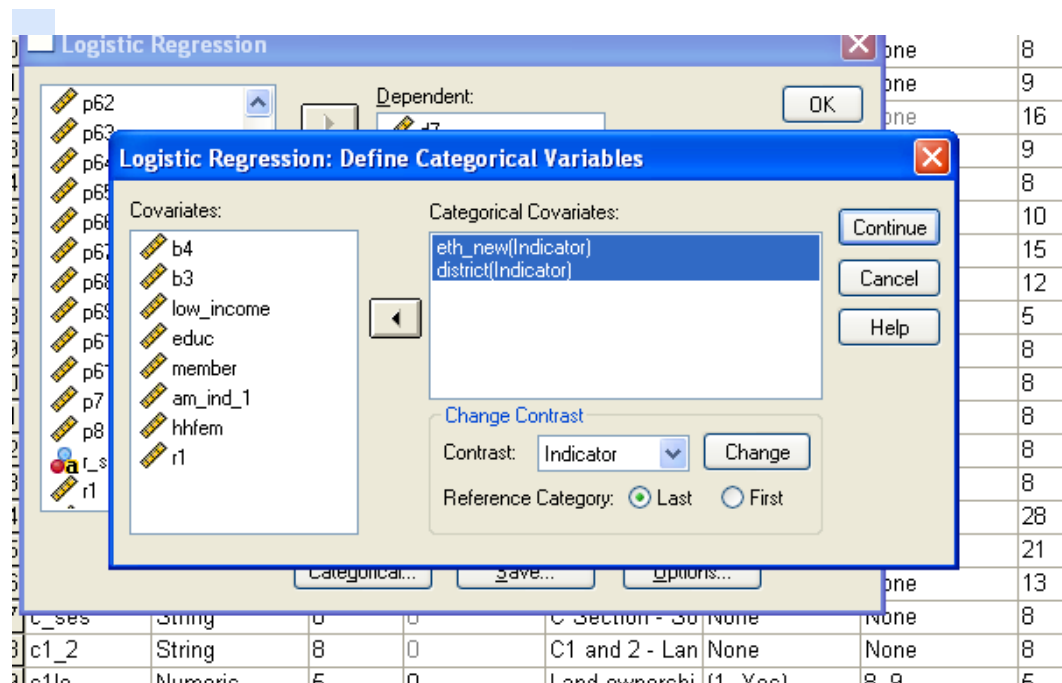
Ethnicity (eth_new)
 District (district)
 Age (b4)
 Sex (b3)
 Poverty (income among 20% lowest (low_income)
 Education (educ)
 Civil society membership (member)
 Household consumer goods level (am_ind_1)
 Female head of household (hh_fem)
 Citizenship (r1)

Perhaps you could think of other variables that should be included?

In the data window, select *Analyze, Regression* and *Binary logistic regression*. Select your dependent variable (job_opp) and your independent variables.



Some of the variables (district, new_eth) are categorical, and need to be defined as such. Click the box Categorical and select these two as categorical:



Default is indicator and last – this means that in your results, the reference categories will be Muslims and Banke, which are those the other categories will be compared with.

Click *Continue* and *OK* (there are many more options, but they will not be explained here).

Let us first take a look at the *Model summary*. It presents two different R square values

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2958,561(a)	,108	,146

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

In the linear regression model (see above), the coefficient of determination, R square, summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables, with larger R square values indicating that more of the variation is explained by the model, to a maximum of 1. For regression models with a categorical dependent variable, it is not possible to compute a single R squared statistic that has all of the characteristics of R square in the linear regression model, so two approximations are computed instead. The following methods are used to estimate the coefficient of determination:

- *Cox and Snell's R square* is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model.
- *Nagelkerke's R square* is an adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1.

What constitutes a “good” R square value varies. These statistics can be suggestive on their own, but they are most useful when comparing competing models for the same data. The model with the largest R squared statistic is “best” according to this measure. In our case, as seen in the table, the R square varies between 0.11 and 0.15.

The *classification table* shows the practical results of using the logistic regression model.

Classification Table ^a					
Observed			Predicted		
			Perceived employment opportunity in government		Percentage Correct
			0 Equal opportunity	1 Less opportunity	
Step 1	Perceived employment opportunity in government	0 Equal opportunity	413	546	43,0
		1 Less opportunity	276	1167	80,9
Overall Percentage					65,8

a. The cut value is ,500

Without knowing the background characteristics of our respondents, if we were to guess their score on the job_opp variable, we would simply guess less opportunity for all respondents, this would be the correct answer in 60% of the cases. However, by knowing the background characteristics on the independent variables, we improve our guess by 6% as shown by the classification table (the Percentage correct is now increased to 65.8%). For each case, the predicted response is Yes if that case's model-predicted probability is greater than the cutoff value specified in the dialogs (in this case, the default of 0.5).

- Cells on the diagonal are correct predictions (413 and 1167).
- Cells off the diagonal are incorrect predictions (276 and 546).

The predictors and coefficient values are used by the procedure to make predictions. The table summarizes the effect of each predictor.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	eth_new			6,059	3	,109	
	eth_new(1)	,093	,214	,192	1	,662	1,098
	eth_new(2)	,455	,239	3,619	1	,057	1,577
	eth_new(3)	,216	,237	,830	1	,362	1,241
	district			122,307	3	,000	
	district(1)	,026	,134	,039	1	,844	1,027
	district(2)	-1,166	,157	54,827	1	,000	,312
	district(3)	-,953	,159	35,851	1	,000	,385
	b4	-,020	,003	37,780	1	,000	,980
	b3	-,223	,102	4,794	1	,029	,800
	low_income	,178	,131	1,860	1	,173	1,195
	educ	-,192	,053	13,203	1	,000	,825
	member	,196	,120	2,645	1	,104	1,216
	am_ind_1	-,212	,031	47,311	1	,000	,809
	hhfem	,588	,206	8,148	1	,004	1,801
	r1	-,029	,119	,060	1	,807	,971
	Constant	2,574	,381	45,702	1	,000	13,123

a. Variable(s) entered on step 1: eth_new, district, b4, b3, low_income, educ, member, am_ind_1, hhfem, r1.

The ratio of the coefficient to its standard error, squared, equals the Wald statistic. If the significance level of the Wald statistic is small (normally less than 0.05, but in our case it has been set to 0.01 due to sampling imperfections) then the parameter is considered useful to the model.

The meaning of a logistic regression coefficient is not as straightforward as that of a linear regression coefficient. While B is convenient for testing the usefulness of predictors, Exp(B) is easier to interpret. Exp(B) represents the ratio-change in the odds of the event of interest for a one-unit change in the predictor. For example, Exp(B) for *educ* is equal to 0.825, which means that the odds of default for a person who has SLC or higher education are 0.825 times the odds of default for a person who has 1-10 grade schooling, which again are 0.825 times the odds of default for a person who is literate but without schooling, and so on, all other things being equal. Values higher than 1 increase the odds, a value lower than 1 decreases the odds.

Let us then interpret our findings:

According to our model the following variables contribute to our model:

District: District is the variable clearly mostly associated with perceived job opportunity. Compared to Banke, people in Sindhupalchok and Surkhet have greater likelihood of perceiving lack of job opportunities, while the situation in Dhanusa is quite similar to that in Banke.

The score on the *consumer goods index* is also very highly associated with the dependent variable: the more access to consumer goods, the less likely a person is to perceive lack of job opportunities. Perception of lack of job opportunities increases with increasing *age*. Education has the opposite effect. Income, citizenship status and

membership in organisations do not contribute much to the model, and should possibly be deleted. It is noteworthy that ethnicity, caste or religious belonging (using our division into four major groups) is not decisive for perception of lack of job opportunities.

As a further check, we can build a model using backward stepwise methods. Backward methods start with a model that includes all of the predictors. At each step, the predictor that contributes the least is removed from the model, until all of the predictors in the model are significant. If the two methods choose the same variables, one can be fairly confident that it's a good model.

8 Presenting your findings – making tables and graphs

How to visualize your findings depends on the purpose of your report or presentation. For an academic audience used to reading tables, this might be a preferred way to present your results. However, in oral presentations with power-point, policy-briefs and papers targeted at a broader audience, a graph very often is easier to interpret, and provides an immediate visual impression of the results.

Here we will only make a few comments on the use of tables.

1. For survey results based on a random selection of respondents and considerable standard errors, it does not make sense to use decimals when presenting percentages of responses. Decimals are slower to read and indicate a greater accuracy than is actually the case.
2. It often makes sense to sort the rows so that the larger numbers stay at the top, unless there are good reasons for not doing so.
3. Usually we put comparisons of interest in vertically.
4. Use a smaller font than you would normally use in the text.
5. Be sure to make a title explaining the table and give enough additional explanation so that it is not necessary to read the text to understand the table.

Let's give an example: We are interested in how often people in the four districts read newspapers. The SPSS raw output gives a table like this:

h2 Listen to Radio * district Survey district Crosstabulation

			district Survey district				Total
			1,00 Dhanusa	2,00 Sindhupalchok	3,00 Surkhet	4,00 Banke	
h2 Listen to Radio	1 All the time	Count	157	56	69	58	340
		% within district Survey district	13,6%	9,7%	11,9%	10,0%	11,8%
	2 Mostly	Count	122	124	120	50	416
		% within district Survey district	10,5%	21,5%	20,8%	8,7%	14,4%
	3 sometimes	Count	14	7	32	1	54
		% within district Survey district	1,2%	1,2%	5,5%	,2%	1,9%
	4 Rarely	Count	215	185	123	156	679
		% within district Survey district	18,6%	32,0%	21,3%	27,0%	23,5%
	5 Not at all	Count	649	206	234	313	1402
		% within district Survey district	56,1%	35,6%	40,5%	54,2%	48,5%
Total		Count	1157	578	578	578	2891
		% within district Survey district	100,0%	100,0%	100,0%	100,0%	100,0%

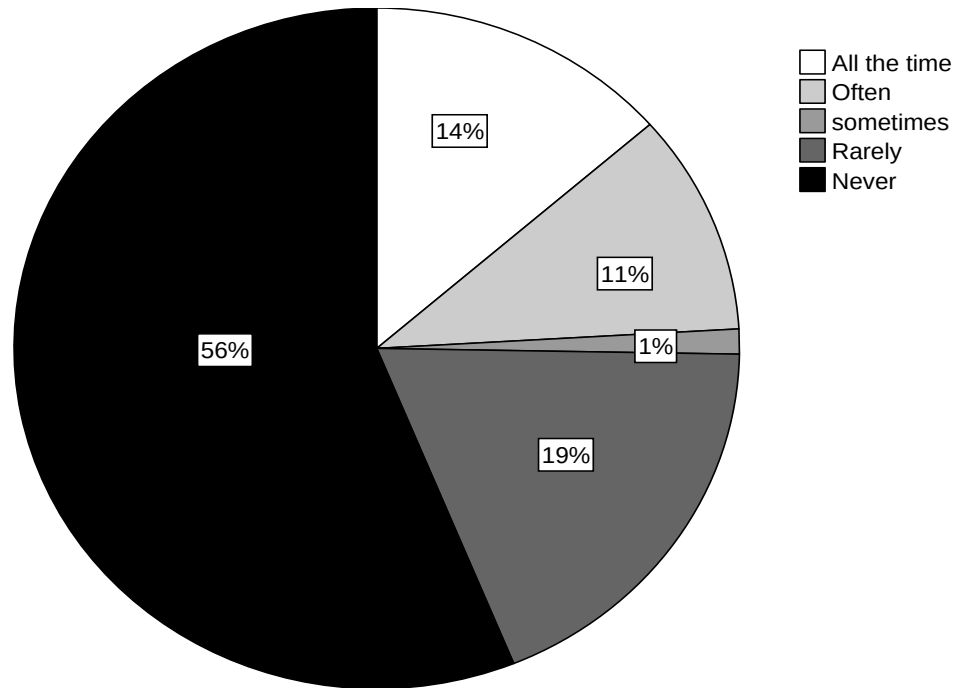
This can be made into a table like this:

Table x.x.: Frequency of listening to radio by district. Percentage of randomly selected respondents (n=2891).

	Dhanusa	Sindhupawlchuk	Surkhet	Banke
Never	56	36	41	54
Rarely	19	32	21	27
Sometimes	1	1	6	0
Often	11	22	21	9
All the time	14	10	12	10
n	1157	578	578	578

When making graphs for univariate distributions, is it better to use a pie chart or a bar chart? The answer is that this depends on the purpose of the chart. Bar charts are usually better if the purpose is to compare individual pieces to each other. Pie charts, on the other hand, are usually better when we wish to compare pieces to the whole.

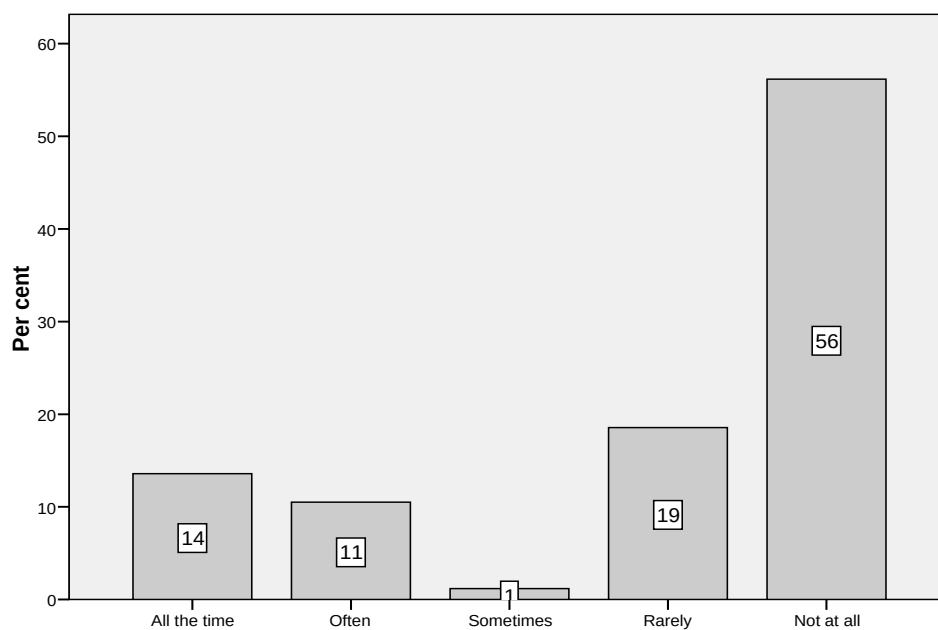
Figure x.x.: Percentage of respondents in Dhanusa with different frequency patterns of listening to radio (n=1157).



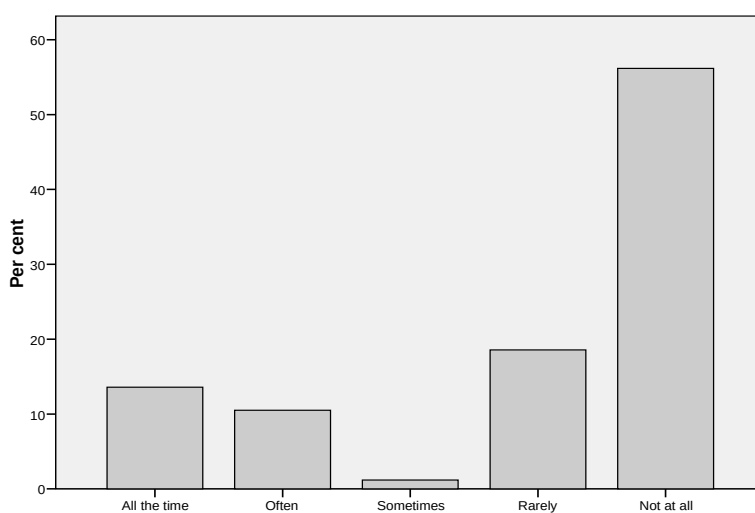
The pie chart is good if we want to see how common the different categories are compared to the total.

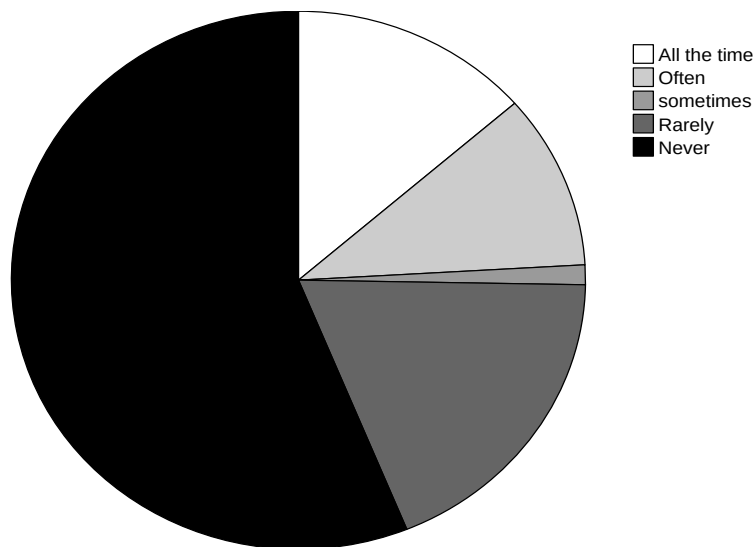
A bar chart would give the following result:

Figure x.x.: Percentage of respondents in Dhanusa with different frequency patterns of listening to the radio (n=1157).



The bar is good if you want to see whether more respondents e.g. answer 'all the time' compared to 'often'. Especially if you don't want to use the labels as in the figures below:





Also, it is recommended to keep the graph simple, and avoid three dimensional and other very fancy graphs, as they tend to be distractive and more difficult to interpret. A good graph relies on simple visual tasks.

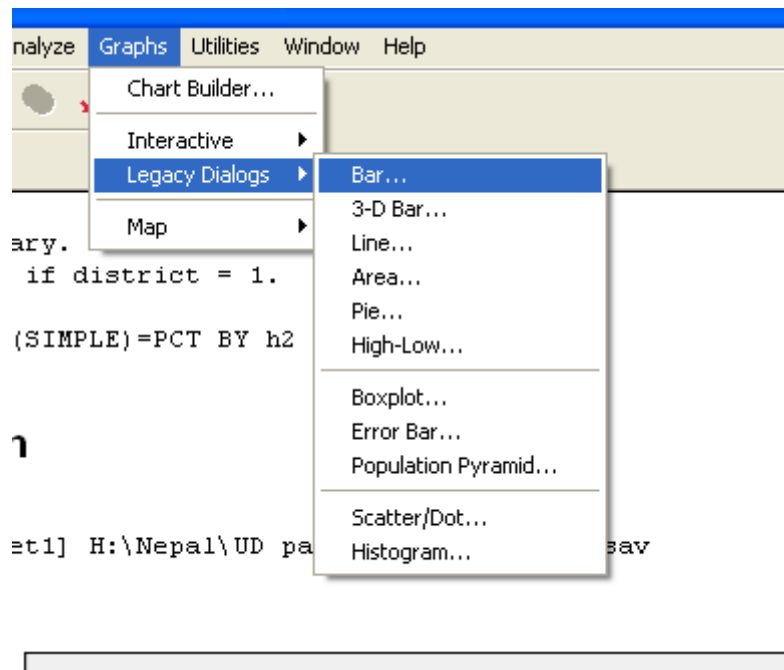
For nominal variables it makes sense to place the bars in order of size. In this way it is easy to see the order of responses. Also, if labels are long, it is easier to fit them into the graph if the barchart is turned sideways.

When we have a number of items represented by different variables, one can use the following procedure to get a good graph:

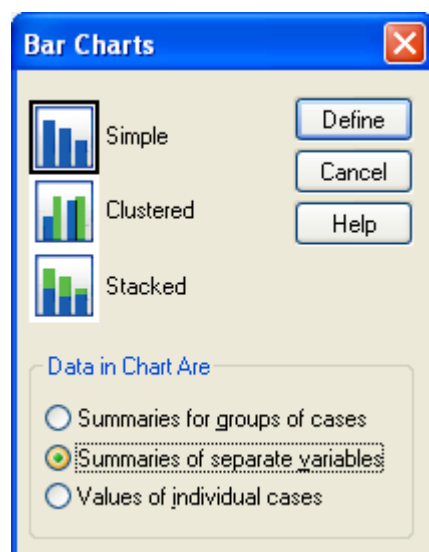
We are interested in the percentage of households in Banke with different types of household consumer items (C20).

First we select only households in Banke. (Select if District = 4).

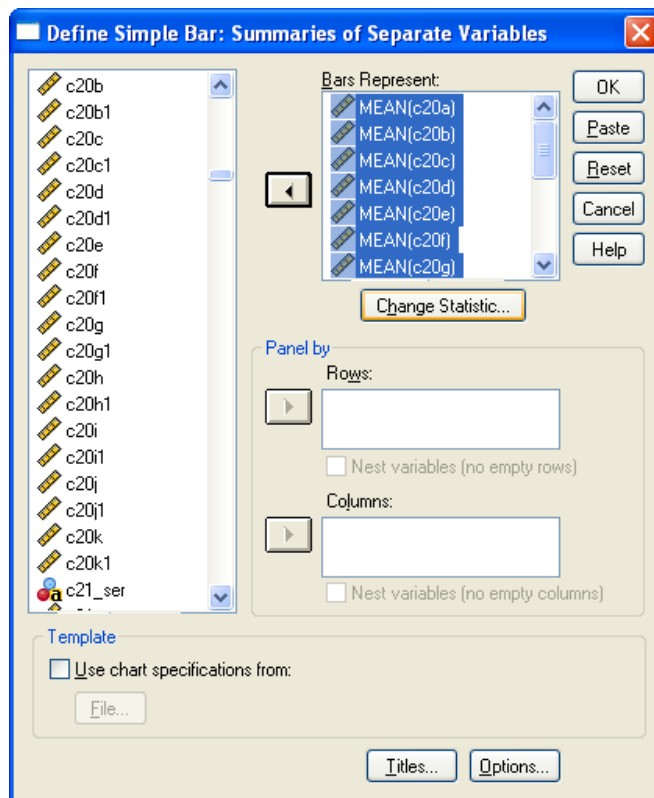
Select Graphs, Legacy dialogues, and Bar...



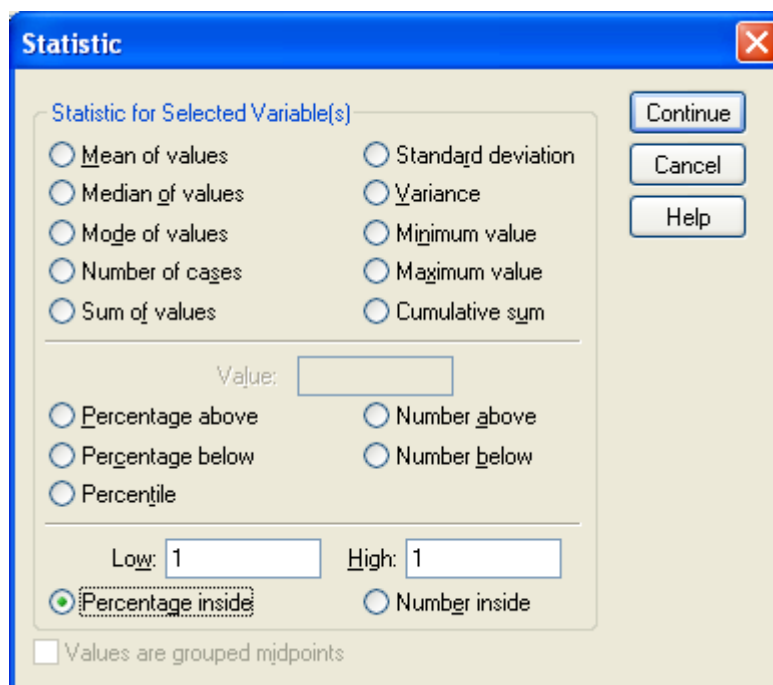
Select Simple (default) and Summaries of separate variables, then Define



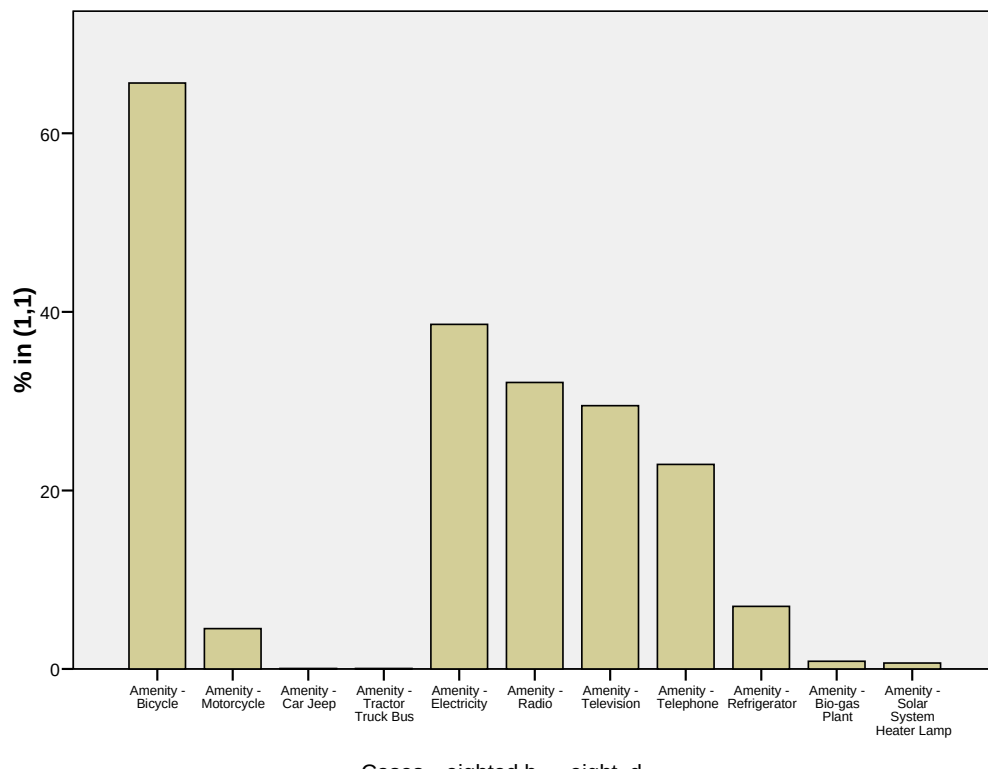
Select C20a to C20k, and press Change statistic



Select percentage inside and fill out Low: 1 High: 1, then Continue

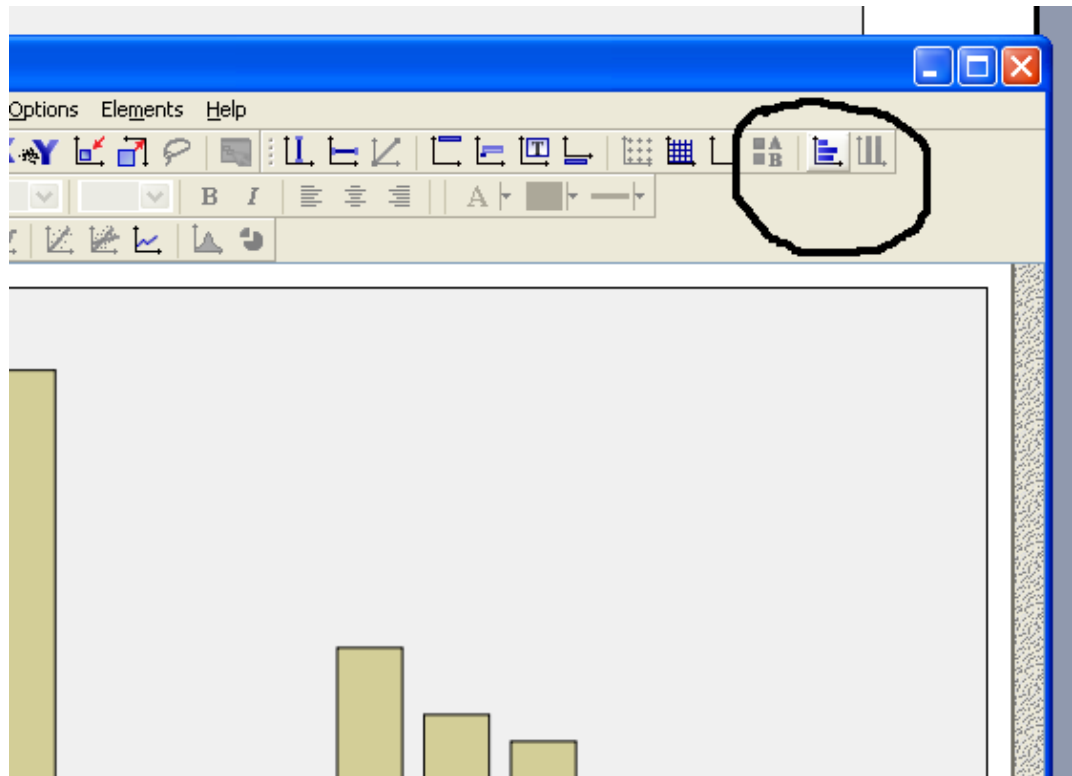


Press OK. Now you will get an overview of all the households with ownership of the listed items:

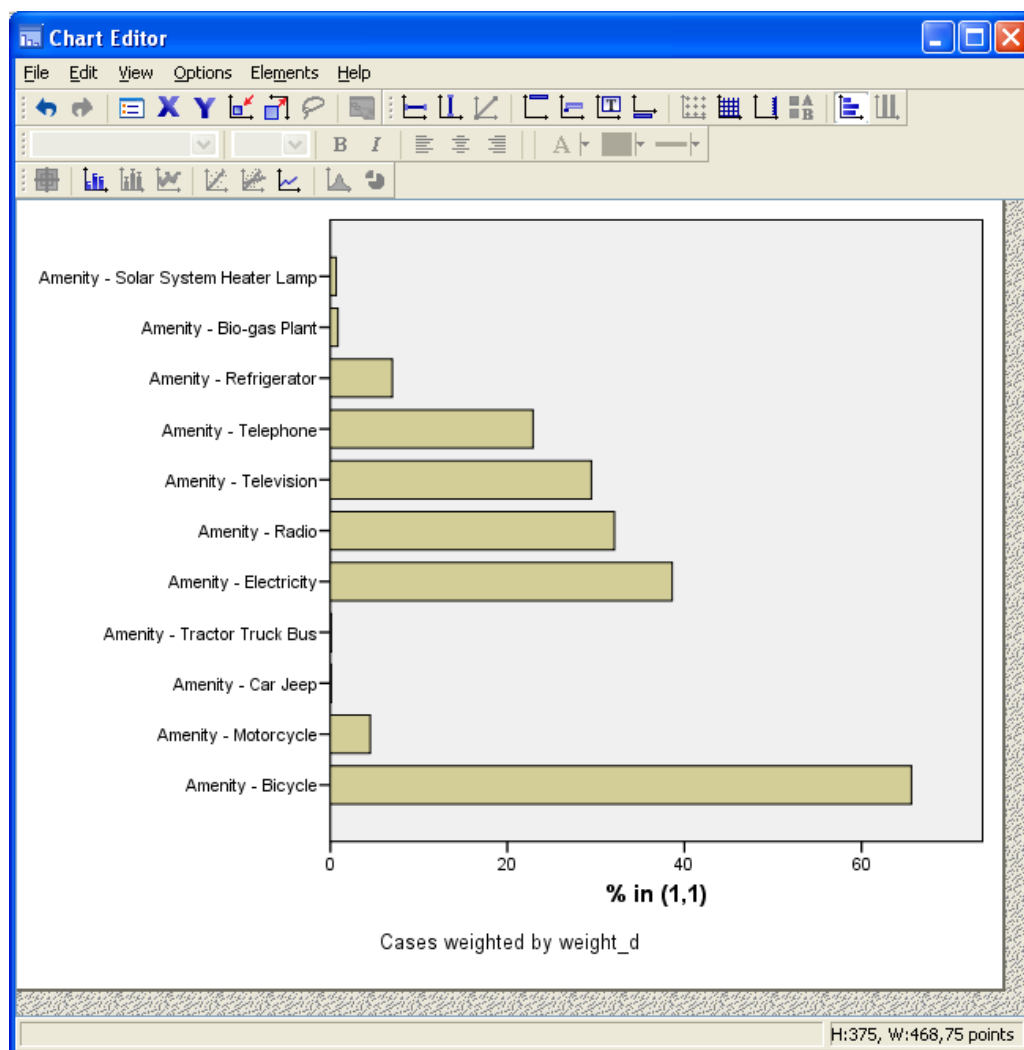


The next steps are a good way to edit the figure. First, we want to turn the graph sideways:

Doubleclick the graph, and start to edit it within the Chart editor window.

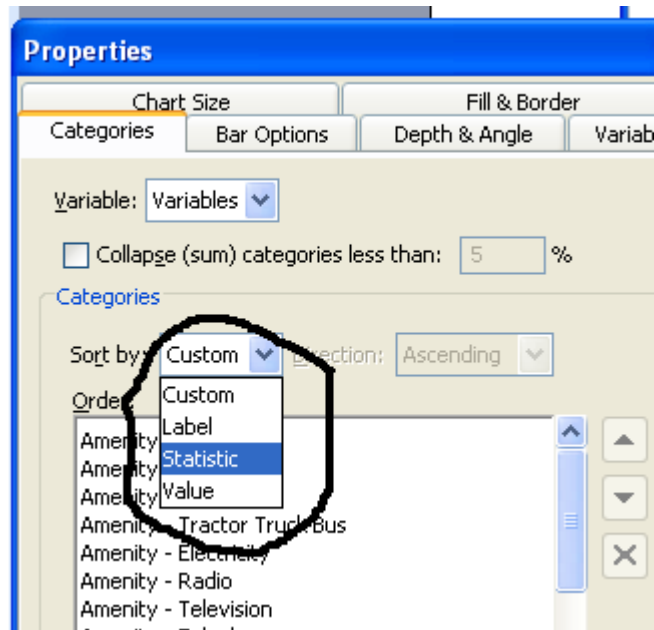


Click the symbol indicated in the above figure (Transpose chart coordinate system).
This gives the following figure:



Now you can start to edit the chart. First you would like to select the order, from high to low:

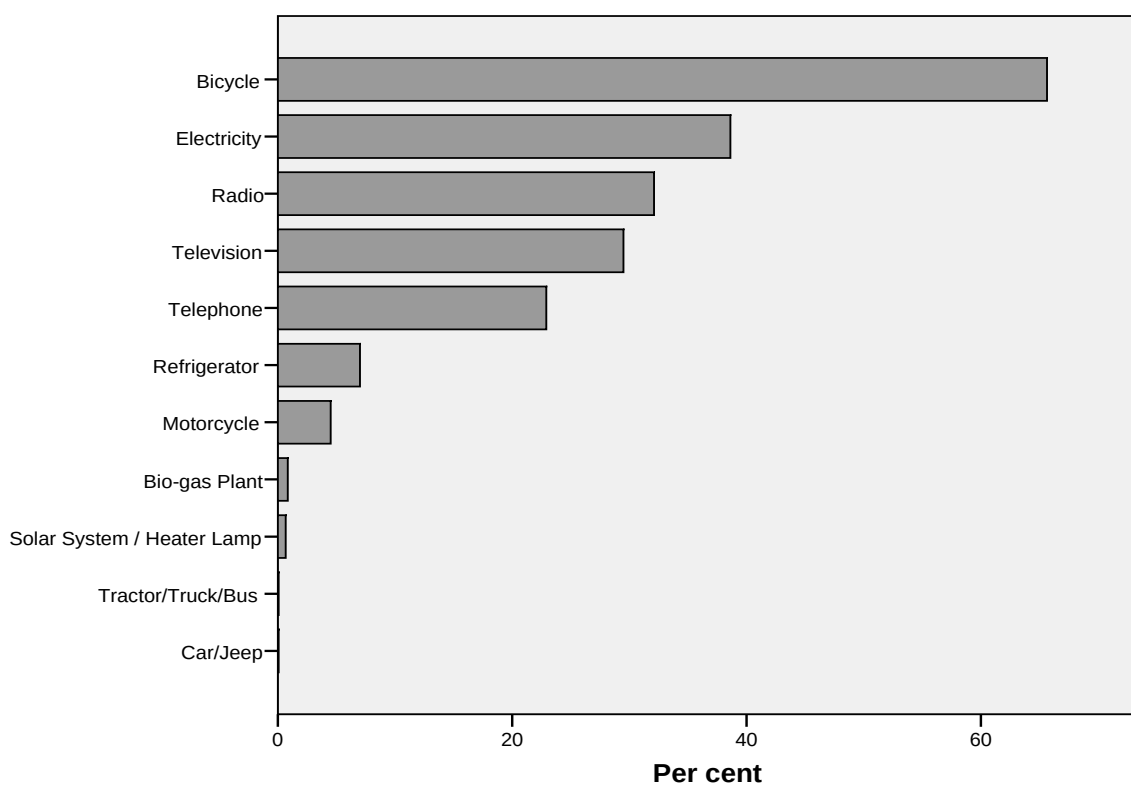
Doubleclick on the bars. The following Properties window appears:



Select Sort by Statistic (either Ascending or Descending according to your taste), and Apply.

After editing some more your chart will look something like this:

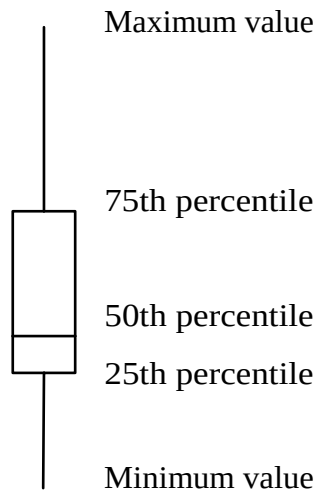
Figure x.x. Percentage of households in Banke with different types of household consumer items.



Additional advice when it comes to making graphs includes the following:

Make different versions of the graph, and choose the one that is best suited. For example, should the graph's axis go from 0 or from somewhere else?

If you have continuous variables and wish to present more than averages (income distribution, etc.), it is sometimes useful to make a box plot. In the box plot you can easily display the maximum and minimum values, the middle of the data, the spread of the data (e.g. 25% and 75% percentiles), and the skewness of the data. See the box plot below for an imagined example:



Be aware of outliers!

Other issues to consider are the *use of colours* (don't use different colours – rather shades - for ordinal data; don't use too bright colours, which may cause optical illusions; don't choose colour combinations that are difficult to distinguish; remember that many people are colour blind), and the *use of symbols* (symbols require use of legend, which may be distractive; more than four symbols tend to overload short term memory; certain symbols – e.g. circles and squares – are easily confused, and especially if they are small).